Delving Deep into Fine-Grained Sketch-Based Image Retrieval

Kaiyue Pang

Submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy*

School of Electronic Engineering and Computer Science

Queen Mary University of London

September 2020

To Mom, Who took me to the library

To Mengwei, Who offers me the kernel of herself

To caffeine and sugar through many a long night

Delving Deep into Fine-Grained Sketch-Based Image Retrieval

Kaiyue Pang

Abstract

To see is to sketch. Since prehistoric times, people use sketch-like petroglyphs as an effective communicative tool which predates the appearance of language tens of thousands of years ago. This is even more true nowadays that with the ubiquitous proliferation of touchscreen devices, sketching is possibly the only rendering mechanism readily available for all to express visual intentions. The intriguing free-hand property of human sketches, however, becomes a major obstacle when practically applied – humans are not faithful artists, the sketches drawn are iconic abstractions of mental images and can quickly fall off the visual manifold of natural objects. When matching discriminatively with their corresponding photos, this problem is known as fine-grained sketch-based image retrieval (FG-SBIR) and has drawn increasing interest due to its potential commercial adoption. This thesis delves deep into FG-SBIR by intuitively analysing the intrinsic unique traits of human sketches and make such understanding importantly leveraged to enhance their links to match with photos under deep learning. More specifically, this thesis investigates and has developed four methods for FG-SBIR as follows:

Chapter 3 describes a discriminative-generative hybrid method to better bridge the domain gap between photo and sketch. Existing FG-SBIR models learn a deep joint embedding space with discriminative losses only to pull matching pairs of photos and sketches close and push mismatched pairs away, thus indirectly align the two domains. To this end, we introduce a generative task of cross-domain image synthesis. Concretely when an input photo is embedded in the joint space, the embedding vector is used as input to a generative model to synthesise the corresponding sketch. This task enforces the learned embedding space to preserve all the domain invariant information that is useful for cross-domain reconstruction, thus explicitly reducing the domain gap as opposed to existing models. Such an approach achieves the first near-human performance on the largest FG-SBIR dataset to date, Sketchy.

Chapter 4 presents a new way of modelling human sketch and shows how such modelling can be integrated into existing FG-SBIR paradigm with promising performance. Instead of modelling the forward sketching pass, we attempt to invert it. We model this inversion by translating iconic free-hand sketches to contours that resemble more geometrically realistic projections of object boundaries and separately factorise out the salient added details. This factorised re-representation makes it possible for more effective sketch-photo matching. Specifically, we propose a novel unsupervised image style transfer model based on enforcing a cyclic embedding consistency constraint. A deep four-way Siamese model is then formulated to importantly utilise the synthesised contours by extracting distinct complementary detail features for FG-SBIR.

Chapter 5 extends the practical applicability of FG-SBIR to work well beyond its training categories. Existing models, while successful, require instance-level pairing within each coarse-grained category as annotated training data, leaving their ability to deal with out-of-sample data unknown. We identify cross-category generalisation for FG-SBIR as a domain generalisation problem and propose the first solution. Our key contribution is a novel unsupervised learning approach to model a universal manifold of prototypical visual sketch traits. This manifold can then be used to paramaterise the learning of a sketch/photo representation. Model adaptation to novel categories then becomes automatic via embedding the novel sketch in the manifold and updating the representation and retrieval function accordingly.

Chapter 6 challenges the ImageNet pre-training that has long been considered crucial by the FG-SBIR community due to the lack of large sketch-photo paired datasets for FG-SBIR training, and propose a self-supervised alternative for representation pre-training. Specifically, we consider the jigsaw puzzle game of recomposing images from shuffled parts. We identify two key facets of jigsaw task design that are required for effective performance. The first is formulating the puzzle in a mixed-modality fashion. Second we show that framing the optimisation as permutation matrix inference via Sinkhorn iterations is more effective than existing classifier instantiation of the Jigsaw idea. We show for the first time that ImageNet classification is unnecessary as a pre-training strategy for FG-SBIR and confirm the efficacy of our jigsaw approach.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged. Some parts of the work have previously been published as:

- **Chapter 3**
  - K. Pang, Y. Song, T. Xiang, T. Hospedales, "Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval", *In Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, 2017.
- **Chapter 4**
  - K. Pang, D. Li, J. Song, Y. Song, T. Xiang, and T. Hospedales, "Deep Factorised Inverse-Sketching", *In Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.
- **Chapter 5**
  - K. Pang*, K. Li*, Y. Yang, T. Hospedales, T. Xiang, and Y. Song, "Generalising Fine-Grained Sketch-Based Image Retrieval", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, 2019.
- **Chapter 6**
  - K. Pang, Y. Yang, T. Hospedales, T. Xiang, and Y. Song, "Solving Mixed-modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, USA, 2020.

Acknowledgements

Song sometimes says, “PhD is not a battle for goals. It is a life achievement with beauty in itself.” Well, after the four-year journey, I am fortunate enough to get some real taste of his words. This would not be possible without the accompany of an incredible group of lovely people, who like gentle winds, have blown me through oceans of bewilderment and despair.

Primarily, I would like to thank, Yi-Zhe Song, for his invaluable guidance, inspiration, enthusiasm, and support throughout the years. Despite mixed feelings about every submission being a last-minute rush, having Song as my PhD advisor is still one of the best decisions of my life. Not only he presented me a picture of the academic world in which I can reside, but also enabled me to appreciate a wider spectrum of life. I think it is particularly worth remembering how we make this: (i) I am possibly Song’s last student that he still gets time to regularly grab some morning coffee with and lies on the grass under the sunshine for non-work banters or general life philosophy discussions. He is too busy for that now and gaining more weight consistently and progressively. (ii) Those days and nights we have both painfully suffered and fervently enjoyed to try to write a sketch idea and always worry it’s not good enough to inspire paid off. We love sketches! There are touching moments as well. I will never forget the morning that Song dropped in specifically to talk and comfort me with four-hour round trip while I was on the verge of a mental breakdown, and his daughter came to the world hours later that day. I first met Song when I was an undergraduate student at his lecture in Beijing. Back then, he said “Kaiyue, I’ve pinned high hope on you. You will be great.” Hope I have become closer to the Kaiyue he once wished. I will always be grateful for his mentorship.

Initialisation matters in deep networks when trained for applicable functionalities. For that, I would like to thank Yanping Zhang for helping to identify my potential chemistry with Song from other advisor options and Yashu Ying for recommending a sophomore highly to Song. Without them, I would not have been on this wonderful journey in the first place. I am also thankful to the very existence of the SketchX lab founded and led by Song. It is a blessing to start a PhD where people around share a similar passion for human sketches. Grateful to all the brainstorming and discussions with peers in SketchX so that I am never a lonely island. These people include Jifei Song, Ke Li, Feng Liu, Qian Yu, Yonggang Qi, Jianshu Zhang, Conghui Hu, Anran Qi, Da Li and Umar Riaz Muhammad.

I had the privilege of conducting and discussing my research with some world-renowned experts over the years. Specially, I would like to thank Tao (Tony) Xiang for saving me out of the "disaster mode" during my early PhD. He always found the positive side of ideas and guided me through many suffering research fruitless periods. Working with Tony has always been pleasant, where he rarely terminates at the big talk with grand vision and motivation, but actively engaging in detailed designs with his insightful observations. Despite being a "big name" already, Tony always arrives at the office earlier than most PhD students and rarely delays a planned meeting or have one without well prepared - he actually keeps a file for every idea ever discussed and unfinished. From him, I witnessed the power of self-rigour and how this sole trait alone can achieve, not to mention combining with a bunch of others. I am grateful to Timothy (Tim) M. Hospedales's help on advising and massaging most of my submissions, where some of them are clearly game-changers. I am constantly impressed by his ability to tease out the nonsense in writing and tell a neat but elegant story out of seemingly complex research ideas. We have spent some wonderful summer afternoons together when attending the same conference. Tim possesses some specific aura that makes time with him always memorable even by just feeling the pass of it. I am looking forward to more of that.

I would like to thank Ioannis (Yiannis) Patras for being my independent assessor throughout my PhD study. My warm appreciation goes to Queen Mary University of London (QMUL) students, staffs and associates for their friendship and various support, in particular (mostly in chronological order): Bizhu Wang, Qi Dong, Melissa Yeo, Tim Kay, Edward Hoskins, Xiatian Zhu, Xun Xu, Li Zhang, Xiaobin Chang, Ryan Layne, Hanxiao Wang, Xu Lan, Wei Li, Elyor Kodirov, Zhiyi Cheng, Ying Zhang, Haris Krikelis, Tianyuan Yu and Jiabo Huang. Special thanks to Miles Hansard for taking me on my PhD program in QMUL when the SketchX transferred to the University of Surrey.

I enjoyed the corridor encounters with Steve Uhlig, where some of them usually developed to longer meets in his office later. Talking with him always gives me new perspectives on things and helps me to think creatively. Steve is also a lively example of a great father. It touches the tender part of me each time he shares the photo of his new-born son with love and pride. For these, I am grateful.

Special thanks to my close collaborators who actually made this thesis possible. In particular, I would like to thank Yongxin Yang for bouncing all my crazy ideas with valuable discussions. Yongxin is always a person of pursue-what-you-love with matching talents and skills and I guess that's the reason behind why he can transform from once a potential banker majoring in accounting and finance to a very successful machine learning researcher now. Afraid I can never be as free as a soul of him, but he inspired me to become a closer one. Thankful to Ke Li, with whom I co-authored my first publication of life when I was still an undergraduate. One lesson I learned going on a trip with Ke is that be always prepared both mentally and physically to the unconventional level of roughness. During his one-year visit to SketchX, we shared lots of wonderful late-night moments and left many unfinished ideas to write. Hope we can finish some of them one day before we get too old.

I am fortunate to meet my cat, Sky. She was a stray back at the time and happened to break into my house to ask for some food. Despite getting along for years, Sky does not fall into the tamed and clingy type and still presents a perfect sense of independence. I learned a lot from this little creature, sometimes by just watching her sleep safe and sound in typical London rainy days. Sky vividly tells me how far our AI is to real intelligence. Possibly I hope AI can never get there.

And finally, I am indebted to my parents and my wife for keeping me fat, fed and happy. They are the realists in my dreamers. I dedicate this thesis to them.

Contents

Chapter 1

Introduction

We are living in a time where visual contents are astoundingly produced and consumed. Inexpensive digital cameras are tirelessly working and form a new dimension of our life – CCTV footage for various hierarchies of security networks, pics and vids we share on social media, and our faces to verify our own phones. The availability of these large-scale visual data and the request of its real-time understanding has enabled researchers to develop powerful and efficient algorithms, which have witnessed impressive progress over the past few years. Given an image, we can now detect the individual objects within (Ren et al., 2015), tell their categories (He et al., 2016) and relationships (Lu et al., 2016), retrieve similar instances (Radenović et al., 2018), and even play a visual-question-answering game (Antol et al., 2015) or translate it to a Monet style painting (Gatys et al., 2016).

While these advances help to better manipulate the visual world around us, most of them are dealing with a single visual domain - natural images in the eyes of cameras. This is not an issue when the visual contents fed to a model can be able directly captured and recorded by the digital devices but is a problem if they only exist in our minds. Indeed, we all ceaselessly perceive information in the visual form, but sadly only a chosen few are skilful enough to effectively express themselves visually. Any less-than-perfect drawing or editing will immediately drive away from visual realism and fall off the visual manifold of natural objects. In other words, we humans are not born as faithful artists. For computer vision problems and learning paradigms that long used to admit data input without a human in the loop, this implies a new family of challenges: can machine handle the human subjectivity in visual rendering?

The recent resurgence of research interests on human free-hand sketching, e.g, sketch recognition (Eitz et al., 2012; Yu et al., 2015), sketch modelling (Riaz Muhammad et al., 2018), sketch-based image retrieval (Sangkloy et al., 2016; Yu et al., 2016) or synthesis (Chen and Hays, 2018; Sangkloy et al., 2017), are attempts to such call. In a broader sense, by explaining seeing via drawing, machine vision is given the best chance to understand how human visual systems operate. Because the domain it functions on offers probably the only visually interpretive way to imitate the imagery processed inside the human brain. In view of the complement to existing computer vision algorithms, the unique traits of sparse black strokes and vast white backgrounds presented in human sketch brings opportunities to re-examine their efficacy and potentially underpin new insights. For example, while deep networks already show superior performance on many perceptual tasks, the ability to quickly adapt or generalise under novel settings has come under increasing scrutiny. The sketch is ideal for learning a meta representation under such purpose as particular styles and details are thrown away to encourage invariance. Models learned on sketches are also less likely to take a tricky shortcut to deceive optimisation objectives, e.g., learning by memorising. The complexity of human-informed data space makes it hard for overfitting unless real understandings of visual primitives are delivered.

This thesis delves deep into an important line of sketch-related research to teach machines to conduct instance-level image search based on human free-hand sketch query input. This is also known as the problem of fine-grained sketch-based image retrieval (FG-SBIR) and has drawn increasing research interests over the past few years due to its potential commercial values. Again, let us be absolutely clear: free-hand sketching is not tracing – there is a fundamental process of abstraction and iconic rendering, where overall geometry is warped and salient details are selectively included. Thus, brute force matching will not solve the problem and deeper computer vision reasoning is necessary to understand the inner workings of sketch-photo matching as humans do.

But why would FG-SBIR be useful? Indeed, compared with mainstream texts-as-query via keywords, attributes and hashtags etc, sketches are intrinsically more flexible and accurate given the rich contents in images and the subjectivity of human perception which is more essential to convey when conducting the search. Visual primitives once hard to be expressed by a common language (as per flexibility) or described quantitatively (as per accuracy) become natural under drawing. The benefits manifest even in our most mundane tasks. Consider the scenario where a lady walks on the street and loves the shoes of a passer-by. How can she look for the shoes on the shopping website later? The conventional way is to type in “a Nike sandal with mid-heel, pointy-toe and slingback” and the search engine will retrieve back a ranked list of images with sorting preference to those tags. However, it is very likely that she will soon find out that there exists noticeable uncertainty in her query itself (e.g., inches of heel, the curvature of the toe and the location of logo.) so that none of the results suits her within acceptable precision. She may also find it difficult trying to enhance the query as that needs words in length of a paragraph to infer both the relative spatial arrangements and stylish subtle details. Conversely, the cumbersome process can be greatly alleviated if replaced by finger sketching her mental picture of that shoes directly on the touchscreen devices. A sketch speaks for a “hundred” words.

In this thesis, we investigate a number of data-driven approaches borrowing the power from deep learning to improve the reliability and practicality of FG-SBIR models. Below we first review previous FG-SBIR works, pinpoint their left unsolved challenges, and demonstrate how we address them in Section 1.1. We then wrap the Introduction part by giving the outline of this thesis in Section 1.2.

1.1 Background, Challenges, and Solutions

Two concurrent works (Sangkloy et al., 2016; Yu et al., 2016) provide the first step towards the capabilities of a practical FG-SBIR system and still underpin the basis of most contemporary FG-

SBIR works. Both models are multi-branch Convolutional Neural Networks (CNNs) designed to learn a joint embedding space in which sketch and photo can be directly compared. The popular pairwise contrastive loss and triplet ranking loss are evaluated by both works and the latter is shown to be superior in both models. The two models differ mainly in whether the photo and sketch CNN branches are Siamese (i.e., with weight sharing) or heterogeneous (i.e., without sharing). Subsequent research focuses on issues surrounding multi-branch deep learning that learns to extract more comparable features. For example, attributes to enable embedding space of enhanced semantics (Li et al., 2017b; Song et al., 2016), attention mechanism for visual focal learning (Song et al., 2017b) and exploration of query synergy between texts and sketches (Song et al., 2017a). It is also noteworthy that unlike some SBIR works (Chen et al., 2009; Zhu et al., 2014) that use a sketch and additional successive steps of text or colour cues to refine retrieval, we cope with non-interactive black & white sketch-based retrieval.

This thesis starts from the prevalent multi-branch deep CNN with a triplet ranking objective and brings forward solutions to the challenges underlying its current practice.

Challenge A: Sketch-photo domain gap There is a large domain gap between sketch and photo – a sketch captures object shape/contour information and contains no information on colour and very little on texture. Existing efforts indirectly narrow the discrepancy between the two domains by learning a discriminative model to pull matching pairs of photos and sketches close and push mis-matched pairs away. Some with additional attempts adopt programmatically extracted edgemaps to either replace photos as input (Yu et al., 2016) or pre-train a matching model with photos (Radenovic et al., 2018) to enhance domain invariance. But the choice to what threshold value we set to eliminate weak and noisy edgemap detection responses from photos remains heuristic.

Solution A: We propose a more principled way to deal with sketch-photo domain gap. The key component is to introduce a generative task of cross-domain image reconstruction – when an input photo is embedded in the joint space, the embedding vector is used as input to a generative model to synthesise the corresponding sketch. Therefore, it explicitly preserves domain-invariant information and reduces the domain gap as opposed to existing models.

Challenge B: Sketch-photo abstraction gap The abstraction issue is manifested in the fact that two people can draw very different sketches of the same object due to different backgrounds, drawing abilities and styles, and different subjective perspectives about the salience of visual elements to include. The resulting gap to match with photos has been tackled together with domain gap via the aforementioned deep metric learning on the whole sketch level. However, a closer inspection of human sketching process reveals that it contains two components with quite distinctive characteristics of abstraction that demand different treatments. The overall geometry of sketch *contours* usually composed of long strokes is heavily warped, while distortion is less of a problem for shorter strokes in *details*. But the choice and amount of details vary by artists.

Solution B: We propose to better bridge the abstraction gap by enabling complementary feature learning between object sketching contours and salient details, i.e., the model needs to learn to separately extract non-overlapping (factorised) features from the two components. This is tackled in two stages. We first show that it is possible to factorise out detail part from a sketch and invert it into a distortion-free object contour via a style transfer model. We then importantly leverage the synthesised contour and feed it together with the original whole sketch to Siamese two deep branches. The complementarity of the outputs is ensured with the decorrelation loss.

Challenge C: Generalisation beyond training categories The promising performance of Siamese triplet network has thus far implicitly assumed the availability of instance-level sketch-photo paired annotations for every coarse category to be evaluated. However, as we find out, it generalises very poorly in practice if training and testing categories are disjoint.

Solution C: We propose a novel framework that automatically adapts the deep feature extraction to a given query sketch. This ensures a good representation is produced at testing-time, even when dealing with out-of-sample data in the form of sketches and photos from new categories. The key component is an auxiliary unsupervised learning approach that maps any given sketch to the manifold embedding that represents a universal dictionary of prototypical sketch traits. The generalisability is then obtained by embedding this universal feature to update the retrieval function accordingly.

Challenge D: Dependence on ImageNet pre-training Compared with datasets for conventional vision task consistently growing bigger towards million scale (Benenson et al., 2019; Carreira et al., 2019; Lin et al., 2014; Rothe et al., 2018), the largest single-category FG-SBIR dataset remains on a size of few thousands. Thus ImageNet pre-training has long been deemed essential by the FG-SBIR community to enable quick competitive performance under deep learning. But the practice of disregarding the fitness between pre-training and the specific downstream task is also intuitively problematic – training for categorisation leads to learning invariance on high-level semantics, while instance level matching asks for both fine-grained and spatially-aware capabilities.

Solution D: We propose a self-supervised pre-training alternative for representation learning. We consider the game of jigsaw puzzle to recover an image from its shuffled patches. By formulating the puzzle in a mixed-modal fashion, a model that can solve it must be able to learn a feature that is: (i) locally invariant to whether a given patch is provided as sketch/photo (since patch modality is randomly selected), and (ii) relationally invariant to the modality of either patch in a disjoint pair (it must be able to use either sketch/photo representation in pairwise comparisons for sorting). It is thus a better-aligned pre-training task for the final task of sketch to photo image retrieval.

1.2 Thesis Outline

This thesis is organised into five chapters:

Chapter 3 presents a discriminative-generative hybrid model for FG-SBIR problem which explicitly aligns the sketch and photo domains. A multi-branch cross-domain deep encoder-decoder model is formulated and in-depth analysis is provided on the model architectural design. The state-of-the-art result is achieved on the largest multi-category FG-SBIR dataset Sketchy (Sangkloy et al., 2016) and to our knowledge the very first study that approaches human performance (50.14% vs. 54.27%).

Chapter 4 identifies the problem of factorised inverse-sketching as a key for both sketch modelling and sketch-photo matching. A novel unsupervised sketch style transfer model is proposed to translate a sketch into a geometrically realistic contour as to invert human sketching process. This makes possible to develop a new FG-SBIR model which separately extracts an object detail representation to complement the synthesised contour for effective matching against photos.

Chapter 5 provides the first solution to the cross-category generalisation problem for FG-SBIR. This is introduced based on a novel universal prototypical visual sketch trait for instance-specific latent domain discovery, which is importantly utilised later to automatically adapt the model embedding for sketches from unseen categories. Extensive experiments validate the efficacy of our method compared to a variety of competitors including direct transfer, other approaches to defining instance-embeddings, and state-of-the-art domain generalisation methods.

Chapter 6 proposes the first study of pre-training approaches for FG-SBIR. A self-supervised objective is formulated to solve the popular game of jigsaw puzzle based on permutation inference via Sinkhorn iterations. Extensive experiments on all four publicly available product-level FG-SBIR datasets show the longstanding practice of ImageNet classification is unnecessary as a pre-training strategy for FG-SBIR and confirm the superiority of our jigsaw approach. The results also show that this leads to improved generalisation across object categories.

Chapter 7 provides a conclusion and suggests several research problems and directions to be pursued as further work.

Chapter 2

Literature Review

This chapter provides a summary of related work to the main contributions of this thesis. We start from an overview for sketch-based image retrieval in Section 2.1, and move onto image-to-image translation works in Section 2.2. Section 2.3 reviews a specific line of methods for domain generalisation and finally self-supervised representation pre-training is covered in Section 2.4.

2.1 Sketch-based Image Retrieval

Research on sketch-based visual search can be traced back to the 1990s, which is of particular focus when the concept of content-based image retrieval was first raised (Hirata and Kato, 1992; Kato et al., 1992). While imagery can be manifested in many forms, e.g., 3D shape, video, the most studied visual search problem based on sketch query is on 2D images, known as the problem of sketch-based image retrieval (SBIR). We will first review some early work relying on hand-crafted features for category-level SBIR, then focus on relevant techniques on deep features and for fine-grained SBIR that go beyond category-level matching precision.

2.1.1 Category-level SBIR

Shallow features Category-level SBIR requires the retrieved photo to come from the same category as the query sketch. A key characteristic of most early works (Cao et al., 2010, 2011; Chans et al., 1997; Del Bimbo and Pala, 1997; Matusiak et al., 1998; Parui and Mittal, 2014; Rajendran and Chang, 2000) is to represent an image as programmatically-generated contour via edge detection (e.g., Canny edge detector) and to match with sketches based on pure local geometric similarity (e.g., blob-based pixels or curvature correlations). More sophisticated hand-crafted features (Bui and Collomosse, 2015; Cao et al., 2013; Parui and Mittal, 2014; Qi et al., 2015; Saavedra et al., 2015; Tolias and Chum, 2017) are later introduced into play with the hope to better bridge the gap between sketches and images. These include explorations of both local (Eitz et al., 2011; Hu and Collomosse, 2013; Hu et al., 2011) and global representations (Chalechale et al., 2004; Eitz et al., 2010; Saavedra, 2014), and histogram of oriented gradients (HOG) (Dalal and Triggs, 2005) is a prevailing choice for both. Local representations are also shown to be superior to its global counterpart across a number of feature types (e.g., scale-invariant feature transform (SIFT) (Lowe, 2004), self-similarity descriptor (SSIM) (Shechtman and Irani, 2007), shape context (SC) (Belongie et al., 2002), HOG) in a comprehensive survey (Hu and Collomosse, 2013). Since each image contains a few thousand of such local descriptors, noisy unfavourable information becomes inevitable. To alleviate this issue, bag-of-words (BoW) model (Eitz et al., 2011) is also explored by first clustering the local features extracted from the whole dataset into k clusters (e.g., k-means) as to build a visual vocabulary. Each local feature of one image is then quantised as a histogram of the weighted distance of visual words. Overall, these shallow features do not achieve semantic understanding and limit their efficacy to simple datasets where sketches within lack of details and only permit exceedingly small local deformations.

Deep features Thanks to the availability of large-scale human sketch datasets (Eitz et al., 2012; Sangkloy et al., 2016), deep convolutional neural networks (CNNs) have been able to apply to SBIR. Sketch-a-Net (Yu et al., 2015) is the first deep CNN model specifically designed for human sketch data, despite it targets on recognition task rather than visual search. (Qi et al., 2016) adapts Sketch-a-Net for SBIR into a two-branch Siamese network by learning a joint embedding space with contrastive loss by pulling similar sketch-photo pairs close and pushing dissimilar ones apart. Such deep metric learning has extended to triplet (Bui et al., 2017) and quardruplet (Seddati et al., 2017) networks and shown promising performance. Additional efforts towards more practical SBIR system are devoted including performance trade-off with the number of training samples and categories (Bui et al., 2018), efficient indexing via short hash code (Liu et al., 2017a) and allowing external aesthetics context to be leveraged in matching (Collomosse et al., 2017). While deep features offering better accuracy compared with learning-free shallow features, their generalisability to novel unseen object categories are usually unsatisfactory and has triggered a relevant line of works on zero-shot SBIR (Dey et al., 2019; Liu et al., 2017b, 2019; Yelamarthi et al., 2018). Methods proposed in this thesis are also based on deep metric learning but step forward to the problem of fine-grained SBIR (FG-SBIR) that requires *instance-level* search precision. Despite FG-SBIR shares many technical insights with aforementioned methods, the data and evaluation change from one-to-many (i.e., a sketch can have many true match gallery photos as long as they come from the same category) to one-to-one (i.e., a sketch normally corresponds *only one* true match gallery photo) potentially invalidates the conclusions from category-level SBIR and makes FG-SBIR a de facto independent field.

2.1.2 Fine-grained SBIR

The problem of fine-grained SBIR was first introduced in (Li et al., 2014) which employs a deformable part-based model (DPM) representation (Felzenszwalb et al., 2009) followed by a graph matching strategy for cross-domain pose correspondence. However, their definition of fine-grain is very different from ours here – a sketch is considered to be a match to a photo if the objects have similar viewpoint, pose and zoom parameters; in other words, they do not necessarily have to contain the same object instance. We attribute the delay of FG-SBIR development to the lack of sketch-photo paired data and expensiveness of collecting them compared to annotating data in more conventional computer vision tasks: the user is first displayed with an image for seconds, and asked to sketch on a blank canvas based on the mental imagery that must reflect the key visual traits essential for instance-level identification. It is until recently FG-SBIR regains the attention of researchers thanks to the advent of three FG-SBIR datasets. Sketchy (Sangkloy et al., 2016) is the largest FG-SBIR dataset to date covering 125 object categories with each category containing 100 natural photos and each photo having at least 5 corresponding human sketches. QMUL-V1 (Yu et al., 2016) focuses on product images and contributes three common fashion categories, namely shoe, chair, and handbag, with each on a size of a few hundreds. Shoe category is later expanded in (Yu et al., 2017b) as QMUL-Shoe-V2 with 6648 sketch-photo pairs, which becomes the current largest single-category FG-SBIR dataset. It is important to note that while Sketchy and QMUL-X datasets are all designed for the task of FG-SBIR, the granularity of their collected sketch-photo pairs is different. This is because the aim of Sketchy is more towards a direct continuation and upgrade of general-purpose category-level SBIR. Gallery photos in Sketchy are natural images carefully curated from ImageNet (Russakovsky et al., 2015), where pose and shape play a noticeable role in differentiating instances within the same category. The sketches rendered based on those photos are thus inevitably affected to overlook or downplay certain key details but still able to conduct an instance-level search from drawer's perspective. On the other hand, the collection of QMUL-X orients from the purpose of commercial applications for shopping and the product photos come with clean background and single pose, and often differ only in diverse stylish parts (e.g., decorations on a shoe body or the buckle type to open a bag). The dataset contributors are also required to work on one category only per time to encourage proficiency. The resulting sketch-photo pairs are understandably more fine-grained where local subtle visual traits must be rendered. The works included in this thesis have developed methods for both datasets.

With the availability of datasets, FG-SBIR is mostly tackled by deep learning. There are mainly two lines of work among the very few studies: (i) How to learn a more comparable sketch-photo joint embedding space. The two pioneering works (Sangkloy et al., 2016; Yu et al., 2016) confirm the efficacy of the paradigm by first pre-training the sketch and photo branches on various perceptual tasks (e.g., ImageNet classification) and fine-tuning with triplet ranking loss on target FG-SBIR dataset. (Song et al., 2017b) improves it by introducing an attention module at conv layer to focus representation learning on specific discriminative local regions rather than being spread evenly over the whole image. Outputs of conv and final fully-connected (FC) layer are further added to keep both coarse and fine semantic information, as similarly proposed in (Yu et al., 2017a). (Song et al., 2016) utilises visual attributes to encourage domain-invariance of the embedding space via integrating extra attribute prediction and attribute-based ranking training objectives. Text is also explored as a complementary input to sketch (Song et al., 2017a) and trained jointly in an end-to-end fashion via two triplet ranking losses - for sketch-photo and text-photo alignment, respectively. (ii) How to train a FG-SBIR model without using any sketch-photo pairs and bypass the needs of the expensive collection process. (Riaz Muhammad et al., 2018) trains a stroke removal policy that learns to predict which strokes can be safely removed without affecting recognisability. This makes possible to transform an edge to multiple images of variable abstractions, which then regard as synthesised sketches to form pseudo sketch-photo pairs for FG-SBIR training. A similar idea is adopted in (Li et al., 2018), which develops a grouper that organise image edges into semantically meaningful parts. The pseudo sketch is then obtained by removing less salient groups with a smaller number of segments, shorter lengths but occupying a bigger region. (Radenovic et al., 2018) takes a different approach by directly assuming edge as a pseudo sketch and train an edge-photo matching network. To reduce the domain gap, an edge filtering layer is introduced to threshold weak edge responses, which typically does not present in human sketching. This thesis falls into the former line of FG-SBIR study and addresses some unique challenges underlying current approaches, as described in Section 1.1.

2.2 Image-to-Image Translation

Remarkable progress has been made on deep generative models, which can be broadly categorised into Variational Autoencoder (VAE) (Kingma and Welling, 2013), Autoregressive Model (Uria et al., 2016) and Generative Adversarial Network (GAN) (Goodfellow et al., 2014). These advances have been actively applied to various practical applications including image stylisation (Gatys et al., 2016; Johnson et al., 2016), single image super-resolution (Ledig et al., 2017), video frame prediction (Mathieu et al., 2016), image manipulation (Korshunova et al., 2017; Zhu et al., 2016) and conditional image generation (Mirza and Osindero, 2014; Odena et al., 2017; Reed et al., 2016; Yan et al., 2016; Zhang et al., 2017). The works most relevant to ours are deep image-to-image translation models (Isola et al., 2017) formulated in conditional generative adversarial network (cGAN). Such a model takes an image as input and produces a bottleneck latent code embedding via an encoder, which is then used by a decoder as input to generate an image in another domain that shares the same identity or semantic information. A naive but common training objective is Euclidean distance between each generated pixel and ground truth counterpart, but this usually causes blurry effect due to its tendency to average all plausible outputs (Mathieu et al., 2016). Therefore, an additional adversarially-trained discriminator (Radford et al., 2015) is introduced in cGAN, where the generated pixels are guided to fool the discriminator from the real ones, and where the discriminator tries best to tell the difference between them. In this way, the blurry pixels will become unacceptable in the eyes of the discriminator and optimise towards sharp visual realism. cGAN also calls for two datasets of training images that represent the visual styles of source and target domain, respectively. The input-output image pairs in a training batch can be either paired or unpaired, where this thesis focus on the latter. Apart from realistic image translation, this thesis also introduces the generative task (as per decoder) as an auxiliary task to help discriminative feature learning (as per encoder). In other words, we do not care about image synthesis quality under this setting. How to improve the encoder is the sole purpose here.

2.2.1 Deep discriminative-generative hybrid models

A desirable property of learning models is the ability to exploit the advantageous information from both discriminative and generative models. A popular early attempt (Bengio et al., 2013) is to first pre-train auto-encoder in a layer-wise fashion via unsupervised reconstruction term and fine-tune the entire stack of encoders in a supervised discriminative manner. Recently, deep discriminative-generative hybrid models have been proposed to leverage both labelled and unlabelled data in a unified framework. An important line of efforts are dedicated to building lateral connections between the symmetric layers of encoder and decoder, thus relieving the pressure of only lower layers are busy at reconstructing while upper layers become idle and not regularised. (Zhao et al., 2015) proposes a stacked what-where autoencoder, where the locations (as per where) of the most numerically activated (as per what) local neurons in an encoding layer are used to guide its corresponding layer in the generative decoder. A similar idea is adopted by (Rasmus et al., 2015). It takes a more direct approach that the output of each decoder layer is determined by both the output of its previous layer and its corresponding encoder layer output via perceptrons. In Chapter 3, we also introduce a generative task and integrate it with FG-SBIR discriminative learning. However, apart from the fact that we are dealing with cross-domain reconstruction between sketch and photo rather than single domain, there is another fundamental difference: instead of trained from scratch, our encoder is already well pre-trained on ImageNet classification task with generic visual understanding learned from million-scale images. The question remains open that given such dramatic imbalance, the addition of generative decoder starting from random parameterisation can even help from the first place. Chapter 3 shows this is possible with asymmetrical design choice, i.e., the complexity of decoder's architecture does not mirror that of the encoder and a different learning strategy to favour the learning of the encoder over the decoder.

2.2.2 Unpaired image-to-image translation

Three concurrent works (Kim et al., 2017; Yi et al., 2017; Zhu et al., 2017) set the basis of how to learn a deep parametric translation function for discovering the underlying relationship between domains with two *independently* collected sets of images and without any extra cross-domain pair labels. Given an image in the source domain, it goes through the encoder-decoder of one cGAN and asked to translate into an image indistinguishable by the discriminator in the target domain. However, the pure supervision on the level of domain leads to potentially many candidates satisfying the condition but failing to pair with the input in a meaningful way. If some paired data are available, the model is at least informed to optimise towards preserving the identity of the input, but without it, an additional structural constraint is expected. The idea of transitivity is thus applied that enforces forward-backward cycle consistency and encourages one-to-one translation. That is, a generated image from one cGAN should be able to reconstruct back to itself via another cGAN. The two cGANs coupled together makes bidirectional mapping between two visual domains possible even under unpaired data setting. In Chapter 4, we aim to stylise a human sketch to a distortion-free contour extracted from the photo edge and frame the problem in the context of unpaired image-to-image translation. And because of the severe perceptual abstraction and line deformation exhibited in sketches, we make two key modifications: (i) The cyclic constraint on visual space is too tight. Instead, we relax it and apply cycle consistency on the embedding space, i.e., given a translated output, it just needs to return to the same place in the high-dimensional embedding with its original input. (ii) Encoder trained from scratch struggles to provide consistently useful gradients for keeping basic visual structures in noisy sketches and sometimes deviates from sane learning to collapse. We replace it with conv layers of VGG-16 (Simonyan and Zisserman, 2015) well pre-trained on ImageNet classification task and keep it fixed throughout. Both qualitative and quantitative results validate the efficacy of both.

2.3 Domain Generalisation via Predicting Novel Domains

Generalising to novel categories beyond the training set is an important capability for computer vision to move out of the lab and impact the real world. This motivates, for example, extensive research in zero-shot object recognition (Changpinyo et al., 2016; Frome et al., 2013; Kodirov et al., 2017). Nevertheless, in the context of SBIR, only two previous works studied cross-category generalisation. Both make use of *external category-level* features to guide learning: (Shen et al., 2018) adopt word-vector of category name learned from language model (Mikolov et al., 2013) to regularise visual learning. By seeing the distributions of words in texts as a semantic space for understanding what objects look like, matching is learned beyond training instances in the visual modality, but also from large, unsupervised text corpus filled with vast amount of human knowledge - which is general. Without ever seeing a cat, such model with visual-semantics embedding may get a bit perplexed with the novel unseen visual trait (e.g., whiskers), but it knows it is closer to the dogs in the training set rather than cars or even tigers as they share similar semantics as small four-leg animals and affinity to humans. Very differently, (Yelamarthi et al., 2018) assumes that ImageNet pre-trained photo features (Simonyan and Zisserman, 2015) are already generalisable enough and focus on bridging sketch-photo heterogeneity by using photo features as guidance for sketch feature regression via a deep conditional generative model. To our best finding, there is no prior work exploring the cross-category generalisation issue in FG-SBIR (CC-FG-SBIR), which is more challenging that requires *generalisable instance-level* differentiation.

Chapter 5 for the first time tackles the problem of CC-FG-SBIR, i.e., a FG-SBIR model generalises well to a novel category without data collection and model re-training. By casting a change of category as domain shift and sketch-photo matching as binary pair classification, solving CC-FG-SBIR is reminiscent of domain generalisation (DG) (Shankar et al., 2018) and domain adaptation (DA) (Csurka, 2017) tasks. Our key idea comes close to a particular group of DG/DA learning approaches: find external descriptors that can improve knowledge sharing across domains and use it to synthesise an appropriate model on the fly for the novel domain (Bertinetto et al., 2016; Lei Ba et al., 2015; Li et al., 2017a; Yang and Hospedales, 2015, 2016). In the context of deep networks, a model is then automatically calibrated where its parameters are predicted from a network conditioned on the external descriptor. Such dynamic parameterisation has been termed hypernetworks (Ha et al., 2017) – where one network synthesises the weights of another. Our proposed method addresses the DG problem in CC-FG-SBIR by embedding the query sketch in a universal embedding space and using this embedding as the meta-descriptor for any sketches coming from the new domain (in place of the external manually-defined descriptor), from which parts of the feature extraction network of both photo and sketch are synthesised (as per hypernetworks).

2.4 Self-supervised Representation Learning

Many deep CNN based computer vision models assume that a rich universal representation has been captured in ImageNet pre-trained CNN (Donahue et al., 2014; Sharif Razavian et al., 2014; Yosinski et al., 2014; Zeiler and Fergus, 2014), which can then be fined-tuned with task-specific data using various strategies (Geng et al., 2016; Long et al., 2015; Ren et al., 2015; Schroff et al., 2015; Xu et al., 2015). Especially for tasks with limited training data, fine-tuning an ImageNet pre-trained model is a near-ubiquitous step, to an extent that its efficacy is rarely questioned. Very recently, (He et al., 2019) challenges the conventional wisdom of ImageNet pre-training for downstream tasks like object detection, and demonstrate how similar results can be obtained by training from scratch. However, even in that study, the scale of data required for effective generalisation is beyond that of typical FG-SBIR datasets used throughout this thesis, thus pre-training for FG-SBIR is a must. In Chapter 6, we show that an appropriately designed self-supervised task (mixed-modal jigsaw solving) and model (permutation inference) leads to a strong initial representation for FG-SBIR that outperforms the classic ImageNet pre-training.

Self-supervised learning is an approach to solving unsupervised learning problems by using the mechanism of supervised learning. The supervision signal is achieved by forming *pretext* tasks from data itself and the learned intermediate representation is expected to carry good semantics or structural understanding that sets as a beneficial initial state to practical *downstream* tasks. Various pretext visual tasks have been proposed, and mostly designed for the downstream purpose of image classification and semantic segmentation. (Doersch et al., 2015) formulates it as predicting the relative position between two random patches from one image. (Noroozi and Favaro, 2016) enhances it to tell the relative positions between all possible combinations of every two parts, i.e., to recover all shuffled patches within an image back to their original locations. A model needs to master the spatial configurations and contexts of objects to play this popular jigsaw game. Another idea is to consider the visual primitives within each patch as a quantifiable attribute vector that can be compared across multiple patches. Simple arithmetic can then be devised for visual learning, e.g., the count of visual primitives within the whole image should be equal to the sum of that in each local patch (Noroozi et al., 2017). Pretext task is also defined on image-level that trains a model to identify the same image with different rotation angles (Gidaris et al., 2018). In a completely different line of work, self-supervised representation learning is regarded as a by-product in generative tasks, e.g., image inpainting (Pathak et al., 2016) or colourisation of greyscale images (Zhang et al., 2016). A key finding of Chapter 6 is what constitutes a “strong” self-supervised method varies dramatically with the downstream task. Methods once work reasonably well for classification may simply fail for FG-SBIR. We also show that Sinkhorn-permutation solution to Jigsaw pre-training is crucial to obtaining dramatic improvement for FG-SBIR as opposed to popular classifier formulation of the problem (Carlucci et al., 2019; Noroozi and Favaro, 2016), where their difference is commonly perceived as minor/negligible before for more conventional vision tasks, e.g., classification/detection (Santa Cruz et al., 2017).

Chapter 3

Cross-domain Generative Training for FG-SBIR

3.1 Background and Motivation

In this chapter, we aim to learn a discriminative-generative hybrid model for FG-SBIR. The state-of-the-art FG-SBIR models (Sangkloy et al., 2016; Yu et al., 2016) are deep models that aim to close the domain gap by learning a joint feature embedding for the two domains. Concretely, multi-branch deep convolutional neural networks (CNNs) are employed where each branch corresponds to one domain and the final shared layer defines the embedding space which is subject to various discriminative losses such as pairwise contrastive loss or triplet ranking loss. These losses are designed to pull matching pairs of photos and sketches close and push mis-matched pairs away. These models thus indirectly align the two domains. However, with limited training data and by focusing only on discriminative losses, these models struggle to capture all the domain-invariant information and thus generalise poorly to test data where the domain discrepancies and misalignments could be different from those in the training data.

Our model also aims to learn a joint embedding space. The key difference to the existing models is that we introduce a generative task of cross-domain image synthesis. When an input photo is embedded in the joint space, the embedding vector is used as input to a generative model to synthesise the corresponding sketch. By doing so, we explicitly enforce the model to preserve all the domain-invariant information in the embedding space. This richer representation thus enables the model to generalise better to unseen test data. More specifically, the proposed model is a multi-branch cross-domain deep encoder-decoder model. The encoder in each branch is a deep CNN that takes an image as input and outputs a feature embedding vector. This vector is then used as input to a deep transposed-convolutional (deconvolutional) network (Zeiler et al., 2010) regularised by the reconstruction loss to reconstruct the corresponding sketch. It is a discriminative-generative hybrid model because both discriminative and generative losses are used for learning the embedding, corresponding to the photo-sketch matching discriminative task and the cross-domain image synthesis generative task respectively.

3.2 Methodology

3.2.1 Network Architecture

Overview The overall network architecture of the proposed discriminative-generative hybrid FG-SBIR model is illustrated in Figure 3.1. It consists of four sub-networks: (1) a three-branch Siamese encoder subnet E that aims to learn a joint embedding space for matching input sketch-photo pairs, (2) a Siamese decoder subnet D that takes an embedding vector and reconstruct a target sketch, (3) a classification subnet C to make the embedding vector class-discriminative and (4) a triplet ranking subnet T to make the vector instance-discriminative. Each encoder branch has the same base network and share their parameters, hence the name Siamese; so does each decoder branch. The four subnets are connected by the joint embedding layer: it is the output of E and input of D, C and T.

Encoder The encoder architecture is based on that of VGGNet (Simonyan and Zisserman, 2015), which has been widely used as the base network in many vision applications. The final classification layer of the network pre-trained on classifying the 1000 ImageNet classes is dropped and an additional shared 256-D fully-connected (FC) layer is added after the 4096-D penultimate FC layer of VGGNet. The ℓ_2 normalised 256-D output of the encoder is the joint

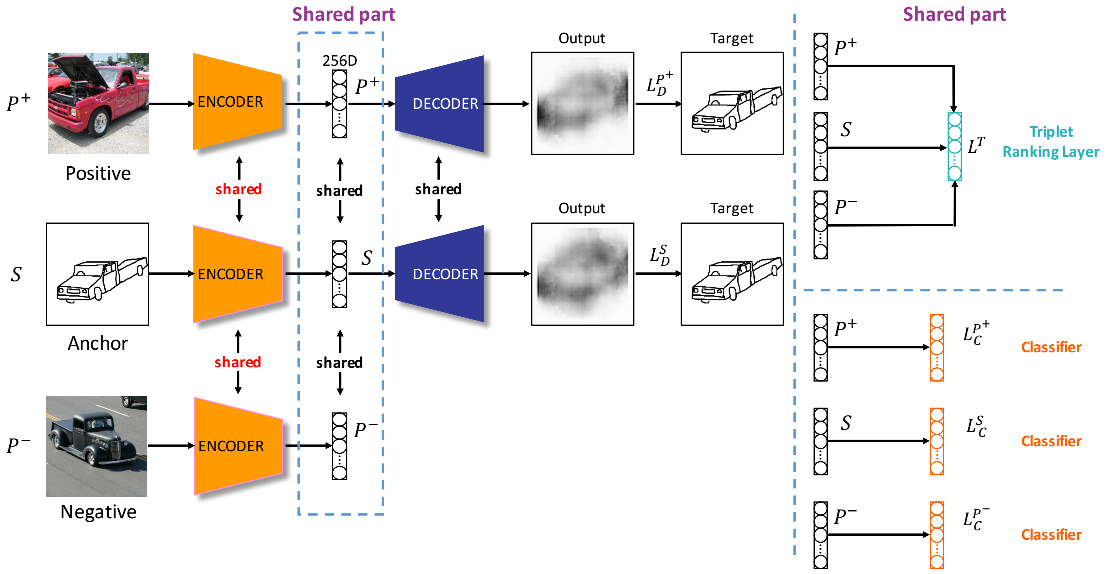


Figure 3.1: Architecture of the proposed deep encoder-decoder FG-SBIR model.

embedding layer and once learned shall be used as the feature representation for both domains for retrieval.

Classification Subnet Although FC layers can be added in the classification subnet, in this work, the classifier directly feeds the latent code to a softmax layer with classification loss L_C being the cross-entropy loss, and the number of output nodes equalling the number of object categories. The classification loss makes sure that the learned embedding space preserves class-discriminative information.

Triplet Ranking Subnet Similar to the classification subnet, we directly add the triplet ranking layer after the shared 256-D embedding layer. In this subnet, each instance tuple $\{s, p^+, p^-\}$ contains an anchor sketch s, a positive photo p^+ containing the same object instance and a negative photo p^-. The subnet has three branches and the goal is to learn a instance-discriminative embedding space where the positive photo p^+ is ranked above the negative photo p^- in terms of its distance to the query sketch s. Note our model is flexible in that any instance-discriminative loss can be used. But as in (Sangkloy et al., 2016; Yu et al., 2016), we found that the triplet ranking loss alone works the best.

Input Size	Filters	Stride	BN	Activation
7 x 7 x 512	512	2 x 2	Yes	ReLU
14 x 14x 512	256	2 x 2	Yes	ReLU
28 x 28 x 256	128	2 x 2	Yes	ReLU
56 x 56 x 128	64	2 x 2	Yes	ReLU
112 x 112 x 64	32	2 x 2	Yes	ReLU
224 x 224 x 32	3	1 x 1	No	Tanh

Table 3-A: Detailed architecture of the decoder subnet.

Decoder The decoder network consists of five upsampling blocks and one final convolution block with a filter size of 4×4 (see Table 3-A for details). Each upsampling block has the structure of Deconvolution-BatchNorm(BN)-ReLU, except the final layer which uses Deconvolution-Tanh for generating the final output. Compared with the encoder-decoder architectures in existing deep generative models (Isola et al., 2017; Sangkloy et al., 2017; Yoo et al., 2016; Zhang et al., 2016), ours differs in that: (i) The decoder is not architecturally symmetric with the encoder. (ii) The decoder is much shallower than the encoder. This design is due to the factor that with the limited training sketch-photo pairs, a deeper decoder network would be prone to overfitting which can make the training process unstable. Furthermore, rather than producing a loyal reconstruction output, the sole objective of this decoder is to help the encoder to learn a richer representation in the embedding layer which is domain-invariant. (iii) The generative process is also asymmetric: We use the embedding vector of the anchor sketch to reconstruct itself, and the positive photo to also reconstruct the anchor sketch. The opposites are not attempted, i.e., sketch-to-photo and photo-to-photo reconstructions. The reason is simple: to compare a photo with a sketch, the additional colour and texture information in the photo domain has to be removed in the embedding layer, so any effort to recover that in the decoder would be futile.

3.2.2 Model Learning and Deployment

Learning Objectives Suppose the encoder, classification and decoder subnets are denoted as ϕ_E, ϕ_C, and ϕ_D, where they are parametrised by θ_E, θ_C and θ_D respectively. Given N sketch-photo triplets $\mathcal{X} = \{\mathbf{x}_i^s, \mathbf{x}_i^{p^+}, \mathbf{x}_i^{p^-}\}_{i=1}^N$ within a training batch, our learning objective is:

$$\underset{\theta_E, \theta_C, \theta_D}{\operatorname{argmin}} \ \mathcal{L} = \mathcal{L}_\mathcal{C} + \lambda_D \mathcal{L}_\mathcal{D} + \lambda_T \mathcal{L}_\mathcal{T}, \tag{3.1}$$

where $\mathcal{L}_\mathcal{C}$ is the cross-entropy softmax loss for classification:

$$L_C = -\sum_{i=1}^{N} (\hat{p_i}^s \log p_i{}^s + \hat{p_i}^{p^+} \log p_i^{p^+} + \hat{p_i}^{p^-} \log p_i^{p^-}), \tag{3.2}$$

$$p_i^{\{s,p^+,p^-\}} = \frac{exp(\phi_C(\phi_E(\mathbf{x}_i^{\{s,p^+,p^-\}})))}{\sum_{j=1}^N exp(\phi_C(\phi_E(\mathbf{x}_j^{\{s,p^+,p^-\}})))}, \tag{3.3}$$

$\mathcal{L}_\mathcal{D}$ is the pixel-wise ℓ_2 reconstruction loss that takes either the input sketch or photo from a ground-truth pair as input and synthesises the input sketch[1] as

$$L_D = \sum_{i=1}^{N} ||\mathbf{x}_i^s - \phi_D(\phi_E(\mathbf{x}_i^s))\|_2 + ||\mathbf{x}_i^s - \phi_D(\phi_E(\mathbf{x}_i^{p^+}))\|_2, \tag{3.4}$$

and $\mathcal{L}_\mathcal{T}$ is the triplet ranking loss:

$$L_T = \sum_{i=1}^{N} \max(0, \Delta + ||\phi_E(\mathbf{x}_i^s) - \phi_E(\mathbf{x}_i^{p^+})\|_2 - ||\phi_E(\mathbf{x}_i^s) - \phi_E(\mathbf{x}_i^{p^-})\|_2). \tag{3.5}$$

λ_D and λ_T weight the three losses by keeping them in roughly the same value range.

Model training strategy The most straightforward way for training a deep model with multiple losses is to update all the parameters together; however the disadvantage of this strategy is that it could lead to detrimental competition between the downstream and upstream tasks. For example, when two sketches belong to different categories but exhibit similar structural and visual cues, θ_E may be sacrificed by pursuing the optimal θ_D. This motivates us an alternate training strategy that learns the encoder first, then fine-tune it with the decoder. One may argue that this would potentially undermine the interpretability of the decoder; nevertheless it is the encoder θ_E that this learning process really cares about, and the image synthesis quality is

[1] We have also experimented adding the popular adversarial loss (Goodfellow et al., 2014) and found that the decoder would suffer from significant mode collapsing problems due to the visual sparsity of sketches, which is commonly observed in generative adversarially trained nets (Salimans et al., 2016).

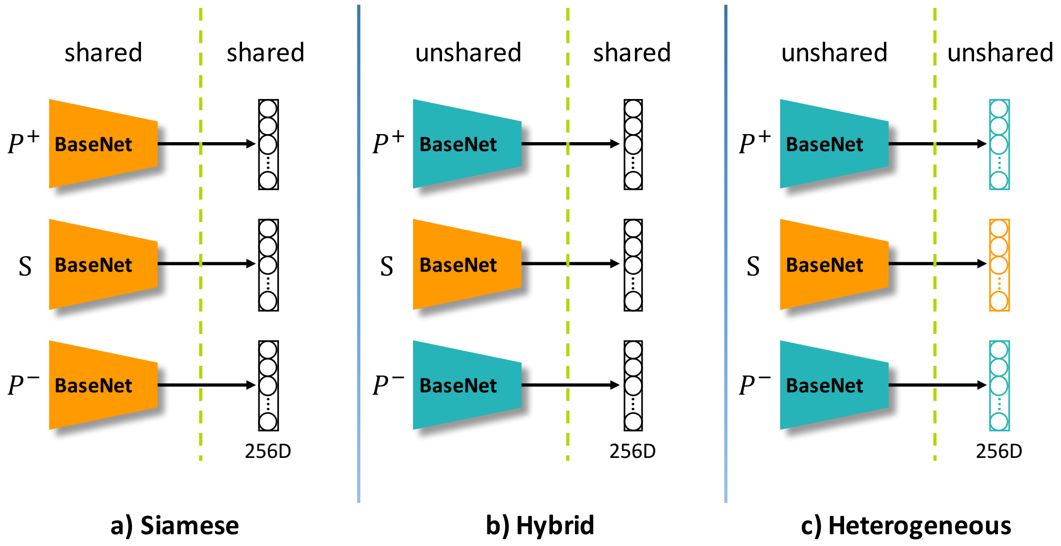


Figure 3.2: Three different weight sharing strategies for FG-SBIR task.

expandable. Specifically, we first minimise $\mathcal{L}_C + \lambda_D \mathcal{L}_D$ with respect to θ_E and θ_C, then minimise $\mathcal{L}$ with respect to θ_E, θ_C and θ_D. In practice, we find this leads to more stable training behaviour.

Model Deployment Once trained, during testing the decoder, classifier and triplet ranking subnets are stripped off. Given a query sketch x^s, we compute the 256D feature representation and use its euclidean distance

$$Dist_{x^s,x^p} = ||\phi_E(\mathbf{x}^s) - \phi_E(\mathbf{x}^p)||_2 \tag{3.6}$$

to rank each photo x^p in the gallery set. Note that we can pre-compute the feature for each photo in the gallery set, which means the retrieval process only involves one forward pass of the encoder followed by Euclidean distance computation; it is thus very efficient.

3.2.3 Discussion on Weight-Sharing Strategies

As illustrated in Figure 3.2, three different weight-sharing strategies exist. Our multi-branch network is Siamese with weight-sharing everywhere between branches. The same strategy is adopted in (Yu et al., 2016). In contrast, the network in (Sangkloy et al., 2016) is heterogeneous meaning there is not weight sharing in any layer between the photo and sketch branches. The Siamese strategy attempts to align the two domains from the very beginning of feature extraction (convolution layers), whilst the heterogeneous one allows feature extraction filters as well as the embedding layer to be learned independently and use the discriminative losses at the network output to align them. As far as domain alignment is concerned, leaving it to the end seems to be counter-intuitive; however, the heterogeneous network has one advantage: one could exploit the far bigger auxiliary data in each domain to pre-train each branch as in (Sangkloy et al., 2016). There is a third way that lies in-between these two extremes: a hybrid strategy whereby the branches are only tied at the joint embedding layer. In our experiments, all three strategies are evaluated.

3.3 Experiments

3.3.1 Experimental Setting

Dataset Experiments are conducted on the Sketchy dataset (Sangkloy et al., 2016), which is the largest free-hand FG-SBIR dataset to date. It contains 125 categories with 100 photos per category and at least 5 sketches for one photo crowd-sourced from Amazon Mechanical Turk (AMT). We use the same training and testing split as in (Sangkloy et al., 2016), where the held-out test set consists of 6312 query sketches and 1250 photos spanning all 125 categories. Another notable FG-SBIR dataset is the QMUL-Shoe-Chair dataset (Yu et al., 2016). However, it is two-magnitudes smaller and contains only two categories. We found that the training of decoder on this dataset is unstable making it difficult to draw any conclusion. It is thus not selected.

Implementation Details Our model is implemented on Tensorflow with a single NVIDIA Tesla P100 GPU. We set the importance weights for different subnets to: $\lambda_D = 10$, $\lambda_T = 1$, with the triplet loss margin $\Delta = 0.1$. The Adam optimiser (Kingma and Ba, 2015) is used with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is set as 10^{-5} at first 20000 iterations and further decreased to 10^{-6} for another 10000 iterations with a batch size of 32. We used the uniformly scaled and centred version of sketches so that the learned representation is not sensitive to the absolute location and scale of a sketch. We randomly cropped an original

256×256 sketch/photo of size 224×224 for data augmentation during training.

Evaluation Metrics We use the same evaluation metrics of recall $@K$ as in (Sangkloy et al., 2016), where for one query sketch, recall $@K$ is 1 if the corresponding photo is within the top K retrieved results and 0 otherwise. We report $acc@1$ by averaging over all queries in the test set.

Competitors To our knowledge, only two works report results on the Sketchy dataset (Liu et al., 2017b; Sangkloy et al., 2016). However, the model in (Liu et al., 2017b) is designed for category-level SBIR with different experiment settings to FG-SBIR, and the focus is on retrieval speed using hashing techniques rather than accuracy. This leaves the various models proposed in (Sangkloy et al., 2016) as the main competitors. These include a heterogeneous GoogLeNet triplet model (**Heter-GN-Tri**), a heterogeneous GoogLeNet pairwise contrastive model (**Heter-GN-Pair**) and a heterogeneous AlexNet pairwise contrastive model (**Heter-AN-Pair**). The other competitor is the Siamese triplet ranking model in (Yu et al., 2016) (**Sia-SN-Tri**). Its base network is called Sketch-a-Net (SN) which is a modified version of AlexNet. It takes an additional preprocessing step to extract edgemaps from photos (Zitnick and Dollar, 2014) in the hope that the domain gap is reduced. For fair comparison, we pre-train the model in stages exactly as described in (Yu et al., 2016) and use the stage-3 pre-trained model to fine tune on the Sketchy dataset with the same classification and triplet ranking losses. The performance of humans (**Human**) on FG-SBIR is also reported in (Sangkloy et al., 2016).

3.3.2 Quantitative Results

Comparisons against the state-of-the-art Our model is compared to the state-of-the-art alternatives as well as humans in Table 3-B. The following observations can be made: (i) Our discriminative-generative hybrid model significantly outperforms all compared models (13.04% improvement over the second best Heter-GN-Tri). (ii) It is now fairly close to the human performance (4.12% lower). (iii) Note that all three heterogeneous baselines in (Sangkloy et al., 2016) took advantage of extensive within-domain pre-training. Our results suggest that it is not necessary with our Siamese hybrid network, significantly simplifying the training process. (iv)

Sia-SN-Tri	**Heter-AN-Pair**	**Heter-GN-Pair**	**Heter-GN-Tri**
16.17%	21.36%	27.36%	37.10%
Ours	**Human** (Sangkloy et al., 2016)		
50.14%	54.27%		

Table 3-B: Comparative results against state-of-the-art FG-SBIR performance.

(a) Contributions of the decoder ϕ_D (vs. -D) and different basenets ϕ_E. (vs. _GN)

Ours-D	**Ours**	**Our_GN-D**	**Ours_GN**
47.18%	**50.14%**	45.52%	48.24%

(b) Contributions of different weight sharing strategies in ϕ_E.

Ours-Heter	**Ours-Hybrid**	**Ours**
41.52%	49.55%	**50.14%**

Table 3-C: Performance of the ablated version of our proposed FG-SBIR model.

The poor result of Sia-SN-Tri (Yu et al., 2016) suggests that replacing natural photos with their edgemap has a negative side-effect given a challenging dataset such as Sketchy. Specifically, as shown in Figure 3.4, the photos in Sketchy often contain other objects and cluttered background. Removing colour information from the very beginning deprives the model of its ability to learn an implicit foreground-background segmentation mechanism to align photos with sketches that have clean background. Note that the objective of the generative decoder is not to synthesise sharp, visually appealing images. Instead, our goal is to reduce the domain gap and extract domain invariant and discriminative features – images in Figure 3.4, albeit blurry, are almost identical when a matching pair of photo and sketch are used as input respectively, showing that this goal has been achieved.

Ablation Study on Competitors Our model differs from competitors in both the base network and the additional generative decoder. To find out what contributes to the superior performance of our model, we compare a few variants with and without the generative decoder and with different base network in Table 3-C(a), where **_GN** refers to replacing our VGGNet with GoogLeNet, **-D** means dropping the generative decoder part. The results show that (i) Regardless of the choice of basenet, adding an additional generative decoder consistently improves the performance and (ii) Compared with GoogLeNet, VGGNet is better for our problem.

Ablation Study on Weight Sharing Strategies In this experiment, we compare our model with the three weight-sharing strategies described in Sec.3.2.3. Table 3-C(b) shows that without any weight-sharing, the heterogeneous version of our model is the weakest in aligning the

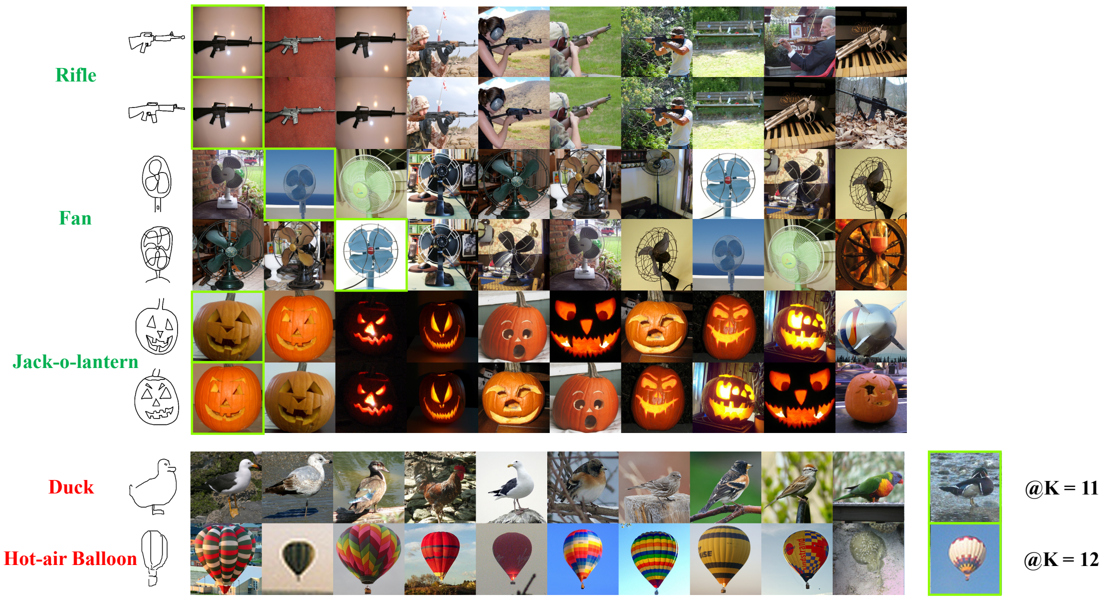


Figure 3.3: Qualitative results. For each query sketch, the top 10 ranked photos out of 1250 candidate photos in the gallery are shown in each row. Green boxes indicate the correct matches and when they are outside the top 10, their actual ranks are given.

two domains, whist the partial-sharing strategy results in a slightly inferior performance. Since a Siamese network has much less parameters than the other two, this justifies the use of the Siamese architecture. Note that such conclusion seemingly counters to some of the existing studies (Bui et al., 2018; Rozantsev et al., 2018) that show partially shared weights can lead to more effective cross-domain image understanding. However, they are not contradictory. Our conclusion suggests that Siamese architecture is a favourable choice when directly fine-tuning a triplet ranking network initialised with ImageNet pre-trained weights for FG-SBIR task. When combined with additional learning signals (e.g., reducing the Maximum Mean Discrepancy between the representations of the two domains as in (Rozantsev et al., 2018)) or pre-training strategies (e.g., multi-stage layer-wise pre-training as in (Bui et al., 2018)), conclusion is understandably subjected to vary.

3.3.3 Qualitative Results

Example retrieval results of the proposed model are shown in Figure 3.3. The results suggest that the model is very effective in removing other objects in the scene and cluttered background

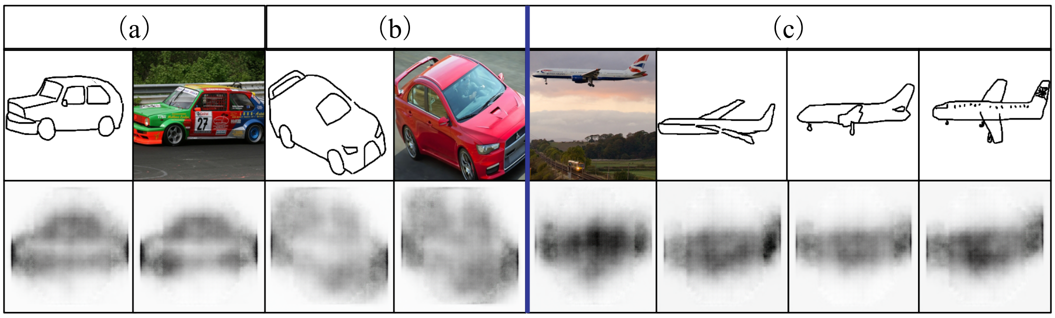


Figure 3.4: Examples of generated images on unseen test sets. In each sub-figure, top: input sketch/photo; bottom: corresponding generated images using the decoder.

and is able to capture subtle instance-level differences, e.g. the first two rifles are matched correctly among some very similar-looking rifle instances. Failure cases are those where true matches are ranked outside the top-10. Two of them are shown in the bottom of Figure 3.3. It is obvious that these failure cases are caused mainly by the poor quality of sketch drawing (e.g., too abstract) with critical details missing in the sketches, giving the model no chance to find the correct matches.

3.3.4 Why Generative Learning Helps

The decoder is designed to help the encoder preserve domain-invariant information. One thus would expect that given a pair of matching photo and sketch, the generated images would be very similar to each other with any domain discrepancies, such as lack of texture, lighting, occlusions and background information in the sketch domain, removed. Figure 3.4 shows that this is exactly what a trained model produces on a test set: (i) Despite the drastically different background clutter (Figure 3.4(a) and (c)) and occlusions (Figure 3.4(b)) exhibited in natural photos, the decoder is able to discard these irrelevant information and focus on the main visual structures. (ii) Given a matched sketch-photo pair, the synthesised images are almost identical; they clearly preserve the shared visual cues (i.e., pose, shape) and neglect the unshared ones such as background and other details ignored by the human sketcher (e.g., Figure 3.4(a), the digit 27 on the side door). (iii) Sketches drawn by different humans for a single photo often varied greatly in the level of abstraction. Figure 3.4(c) shows that our decoder normalises these variations making the retrieval task easier. We thus conclude that having the generative decoder encourages the learned feature representation in the joint embedding layer to focus on the cross-domain shared semantic visual cues rather than the domain-specific information. Importantly it directly reduces the domain gap and enables the learning of a richer representation useful for model generalisation.

3.4 Summary

This chapter has proposed a hybrid discriminative-generative approach for FG-SBIR based on a cross-domain deep encoder-decoder network architecture. The hypothesis was that by introducing the additional generative task, the learned joint embedding space would capture domain-invariant information and explicitly reduce the domain gap between photo and sketch. Extensive experiments validated the hypothesis and demonstrated that the proposed model outperforms existing discriminative models by a large margin.

Chapter 4

Deep Factorised Inverse-Sketching for FG-SBIR

4.1 Background and Motivation

In this chapter, we aim to devise a framework for inverting the iconic rendering process in human free-hand sketch, and for contour-detail factorisation learning in FG-SBIR. A closer inspection of the human sketching process reveals that it includes two components. As shown in (Li et al., 2017c), a sketcher typically first deploys long strokes to draw iconic object contours, followed by shorter strokes to depict visual details (e.g., shoes laces or buckles in Figure 4.1(a)). Both the iconic contour and object details are important for recognising the object instance and matching a sketch with its corresponding photo. The contour is informative about object subcategory (e.g., a boot or trainer), while the details distinguish instances within the subcategory – modelling both are thus necessary. However, they have very different characteristics demanding different treatments. The overall geometry of the sketch contour experiences large and user-specific distortion compared to the true edge contour of the photo (compare sketch contour in Figure 4.1(a) with photo object contour in Figure 4.1(b)). Photo edge contours are an exact perspective projection of the object boundary; and free-hand sketches are typically an orthogonal projection at best, and usually much more distorted than that – if only because humans seem unable to draw long

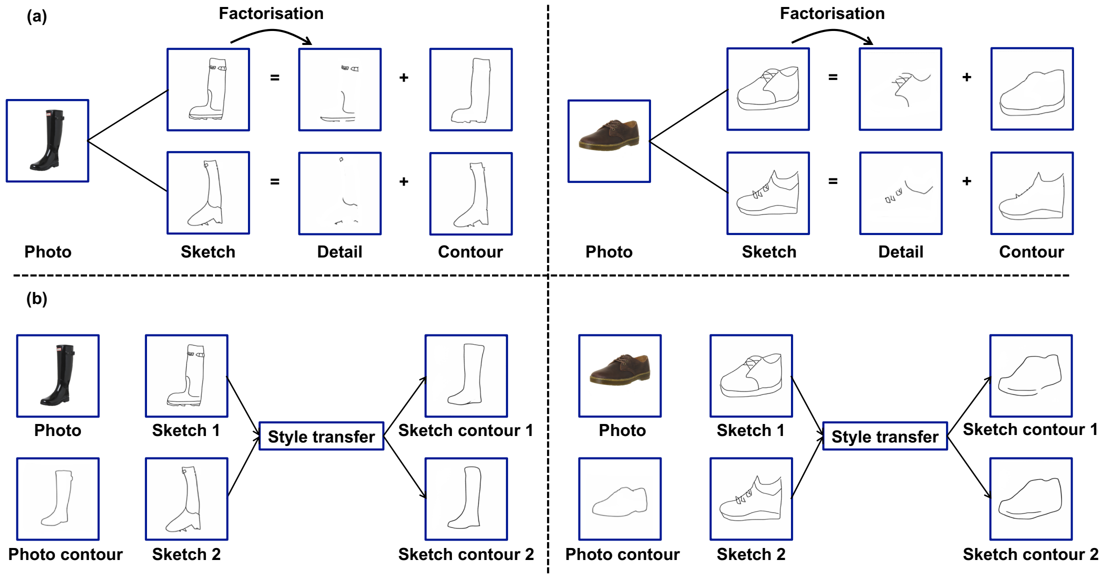


Figure 4.1: (a) A free-hand object instance sketch consists of two parts: iconic contour and object details. (b) Given a sketch, our style transfer model restyles it into distortion-free contour. The synthesised contours of different sketches of the same object instance resemble each other as well as the corresponding photo contour.

smooth lines without distortion (Flash and Hogan, 1985). In contrast, distortion is less of an issue for shorter strokes in the object detail part. But choice and amount of details varies by artist (e.g., buckles in Figure 4.1(a)).

We propose to model human sketches by inverting the sketching process. That is, instead of modelling the forward sketching pass (i.e., from photo/recollection to sketch), we study the inverse problem of translating sketches into visual representations that closely resemble the perspective geometry of photos. We further argue that this inversion problem is best tackled on two levels by separately factorising out object contours and the salient sketching details. Such factorisation is important for both modelling sketches and matching them with photos. This is due to the differences mentioned above: sketch contours are consistently present but suffer from large distortions, while details are less distorted but more inconsistent in their presence and abstraction level. Both parts can thus only be modelled effectively when they are factorised.

We tackle the first level of inverse-sketching by proposing a novel deep image synthesis model for style transfer. It takes a sketch as input, restyles the sketch into natural contours resembling the more geometrically realistic contours extracted from photo images, while removing object details (see Figure 4.1(b)). This stylisation task is extremely difficult because (a) Collecting a large quantity of sketch-photo pairs is infeasible so the model needs to be trained in an unsupervised manner. (b) There is no pixel-to-pixel correspondence between the distorted sketch contour and realistic photo contour, making models that rely on direct pixel correspondence such as (Isola et al., 2017) unsuitable. To overcome these problems, we introduce a new cyclic embedding consistency in the proposed unsupervised image synthesis model. It forces the sketch and unpaired photo contours to share some support in a common low-dimensional semantic embedding space.

We next complete the inversion in a discriminative model designed for matching sketches with photos. It importantly utilises the synthesised contours to factor out object details to better assist with sketch-photo matching. Specifically, given a training set of sketches, their synthesised geometrically-realistic contours, and corresponding photo images, we develop a new FG-SBIR model that extracts factorised feature representations corresponding to the contour and detail parts respectively before fusing them to match against the photo. The model is a deep Siamese neural network with four branches. The sketch and its synthesised contours have their own branches respectively. A decorrelation loss is applied to ensure the two branch's representations are complementary and non-overlapping (i.e., factorised). The two features are then fused and subject to triplet matching loss with the features extracted from the positive and negative photo branches to make them discriminative.

4.2 Sketch Stylisation with Cyclic Embedding Consistency

Problem definition: Suppose we have a set of free-hand sketches S drawn by amateurs based on their mental recollection of object instances (Yu et al., 2016) and a set of photo object contours C sparsely extracted from photos using an off-the-shelf edge detection model (Zitnick and Dollar, 2014), with empirical distribution $s \sim p_{data}(S)$ and $c \sim p_{data}(C)$ respectively. They are thematically aligned but otherwise *unpaired* and *non-overlapped* meaning they can contain different sets of object instances. This makes training data collection much easier. Our objective is to learn an unsupervised deep style transfer model, which inverts the style of a sketch to a cleanly rendered object contour with more realistic geometry, and user-specific details removed (see Figure 4.1(b)).

4.2.1 Model Formulation

Our model aims to transfer images in a source domain (original human sketches) to a target domain (photo contours). It consists of two encoder-decoders, $\{E_S, G_S\}$ and $\{E_C, G_C\}$, which map an image from the source (target) domain to the target (source) domain and produce an image whose style is indistinguishable from that in the target (source) domain. Once learned, we can use $\{E_S, G_C\}$ to transfer the style of S into that of C, i.e., distortion-free and geometrically realistic contours. Note that under the unsupervised (unpaired) setting, such a mapping is highly under-constrained – there are infinitely many mappings $\{E_S, G_C\}$ that will induce the same distribution over contours c. This issue calls for adding more structural constraints into the loop, to ensure s and c lie on some shared embedding space for effective style transfer and instance identity preserving between the two. To this end, the decoder G_S (G_C) is decomposed into two sub-networks: a shared embedding space construction subnet G_H, and an unshared embedding decoder $G_{H,S}$ ($G_{H,C}$), i.e., $G_S \equiv G_H \circ G_{H,S}, G_C \equiv G_H \circ G_{H,C}$ (see Figure 4.2(a)).

Embedding space construction: We construct our embedding space similarly to (Liu and Tuzel, 2016; Liu et al., 2017c): The G_H projects the outputs of the encoders into a shared embedding space. We thus have $h_s = G_H(E_S(s)), h_c = G_H(E_C(c))$. The projections in the embedding space are then used as inputs by the decoder to perform reconstruction: $\hat{s} = G_{H,S}(h_s), \hat{c} = G_{H,C}(h_c)$.

Embedding regularisation: As illustrated in Figure 4.2 (b), the embedding space is learned with two regularisations: (i) Cyclic embedding consistency: this exploits the property that the learned style transfer should be 'embedding consistent', that is, given a translated image, we can arrive at the same spot in the shared embedding space with its original input. This regularisation is formulated as:

$$h_s = G_H(E_S(s)) \rightarrow G_{H,C}(G_H(E_S(s))) \rightarrow G_H(E_C(G_{H,C}(G_H(E_S(s))))) \approx h_s$$

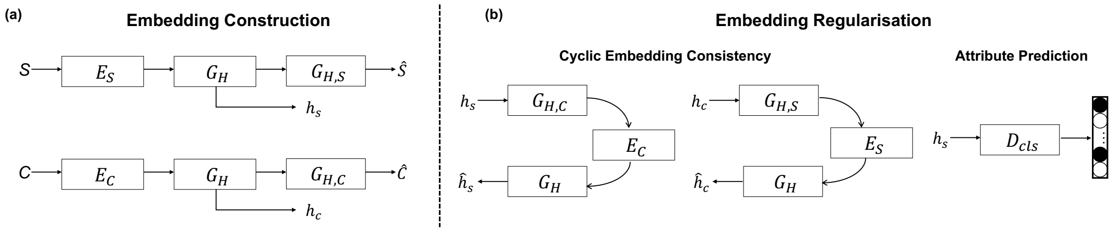


Figure 4.2: Schematic of our sketch style transfer model with cyclic embedding consistency. (a) Embedding space construction. (b) Embedding regularisation via cyclic embedding consistency and an attribute prediction task. $E_S, G_{H,S}$ and $E_C, G_{H,C}$ represent domain-specific encoder-decoder for sketch and contour respectively. G_H is the shared domain embedding learner.

$$h_c = G_H(E_C(c)) \rightarrow G_{H,S}(G_H(E_C(c))) \rightarrow G_H(E_S(G_{H,S}(G_H(E_C(c))))) \approx h_c$$

for the two domains respectively. This is different from the cyclic visual consistency used by existing unsupervised image-to-image translation models(Liu and Tuzel, 2016; Liu et al., 2017c; Zhu et al., 2017), by which the input image is reconstructed by translating back the translated input image. The proposed cyclic embedding consistency is much ‘softer’ compared to the cyclic visual consistency since the reconstruction is performed in the embedding space rather than at the per-pixel level in the image space. It is thus more capable of coping with domain discrepancies caused by the large pixel-level mis-alignments due to contour distortion and the missing of details inside the contours. (ii) Attribute prediction: to cope with the large variations of sketch appearance when the same object instance is drawn by different sketchers (see Figure 4.1(a)), we add an attribute prediction task to the embedding subnet so that the embedding space needs to preserve all the information required to predict a set of semantic attributes.

Adversarial training: Finally, as in most existing deep image synthesis models, we introduce a discriminative network to perform adversarial training (Goodfellow et al., 2014): the discriminator is trained to be unable to distinguish generated contours from sketch inputs and the photo contours extracted from object photos.

4.2.2 Model Architecture

Encoder: Most existing unsupervised image-to-image translation models design a specific encoder architecture and train the encoder from scratch. We found that this works poorly for

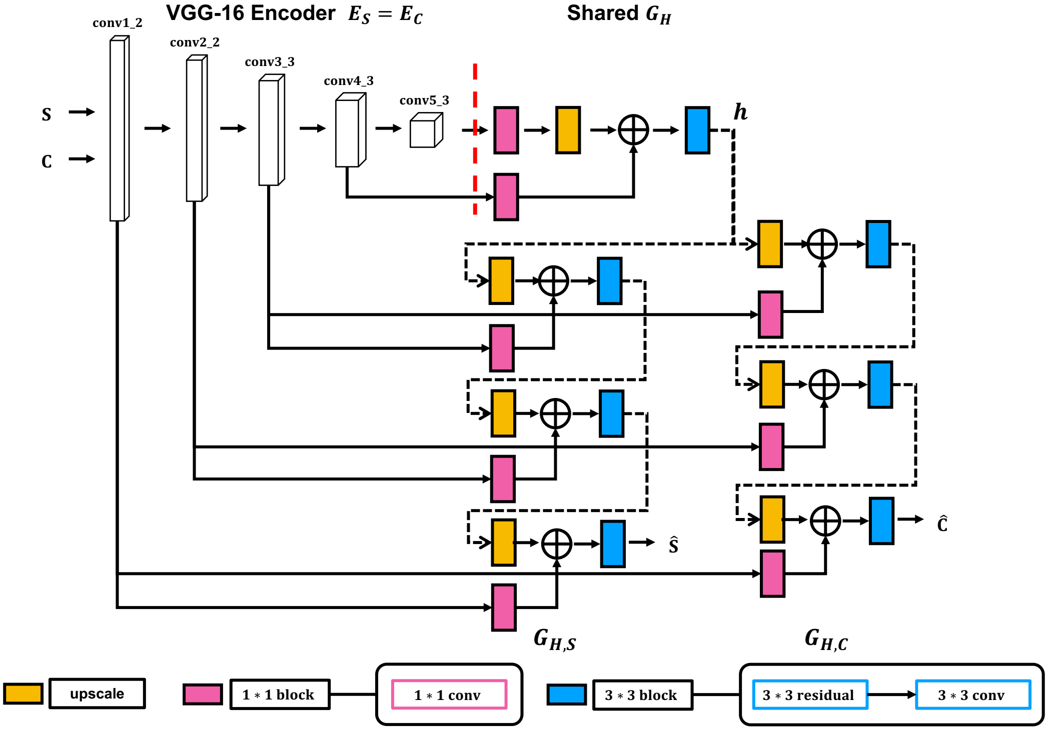


Figure 4.3: A schematic of our specifically-designed encoder-decoder.

sketches due to lack of training data and the large appearance variations mentioned earlier. We therefore adopt a fixed VGG encoder pre-trained on ImageNet. This is particularly critical at the beginning of learning to ensure the basic visual structures presented in both s and c to be retained while otherwise prone to lost and derailed to trivial solutions. As shown in Figure 4.3, the encoder consists of five convolutional layers before each of the five max-pooling operations of a pre-trained VGG-16 network, namely $conv1\_2$, $conv2\_2$, $conv3\_3$, $conv4\_3$ and $conv5\_3$. Note that adopting a pre-trained encoder means that now we have $E_S = E_C$.

Decoder: The two subnets of the decoder: G_H and $G_{H,S}$ ($G_{H,C}$) use a residual design. Specifically, for convolutional feature map extracted at each spatial resolution, we start with $1 * 1$ conv, upsample it by a factor of 2 with bilinear interpolation and then add the output of the corresponding encoder layer. It is further followed by a $3 * 3$ residual and $3 * 3$ conv for transformation learning and adjusting appropriate channel numbers for the next resolution. Note that shortcut connections between the encoder and decoder corresponding layers are also established in the residual form. As illustrated in Figure 4.3, the shared embedding construction subnet G_H is composed of one such block while the unshared embedding decoders $G_{H,S}$ ($G_{H,C}$) have three. For more details of the encoder/decoder and discriminator architecture, please see Sec. 4.4.1.

4.2.3 Learning Objectives

Embedding consistency loss: Given s (c), and its cross-domain synthesised image $G_C(E_S(s))$ ($G_S(E_C(c))$), they should arrive back to the same location in the embedding space. We enforce this by minimising the Euclidean distance between them in the embedding space:

$$\begin{aligned}\mathcal{L}_{embed} = \mathbb{E}_{s\sim S,c\sim C}[||G_H(E_S(s)) - G_H(E_C(G_C(E_S(s))))||_2 \\ +||G_H(E_C(c)) - G_H(E_S(G_S(E_C(c))))||_2].\end{aligned} \tag{4.1}$$

Self-reconstruction loss: Given s (c), and· its reconstructed result $G_S(E_S(s))$ ($G_C(E_C(c))$), they should be visually close. We thus have

$$\mathcal{L}_{recons} = \mathbb{E}_{s\sim S,c\sim C}[||s - G_S(E_S(s))||_1 + ||c - G_C(E_C(c))||_1]. \tag{4.2}$$

Attribute prediction loss: Given a sketch s and its semantic attribute vector a, we hope its embedding $G_H(E_S(s))$ can be used to predict the attributes a. To realise this, we introduce an auxiliary one-layer subnet D_{cls} on top of the embedding space h and minimise the classification errors:

$$\mathcal{L}_{cls} = \mathbb{E}_{s,a\sim S}[-\log D_{cls}(a|G_H(E_S(s)))]. \tag{4.3}$$

Domain-adversarial loss: Given s (c) and its cross-domain synthesised image $G_C(E_S(s))$ ($G_S(E_C(c))$), the synthesised image should be indistinguishable to a target domain image c (s) using the adversarially-learned discriminator, denoted D_C (D_S). To stabilise training and improve the quality of the synthesised images, we adopt the least square generative adversarial network (LSGAN) (Mao et al., 2017) with gradient penalty (Gulrajani et al., 2017). The domain-adversarial loss for generator is defined as:

$$\begin{aligned}\mathcal{L}_{adv_g} &= \mathbb{E}_{s\sim S}[||D_C(G_C(E_S(s))) - 1||_2] \\ &+ \mathbb{E}_{c\sim C}[||D_S(G_S(E_C(c))) - 1||_2]\end{aligned} \tag{4.4}$$

and for the discriminator:

$$\begin{aligned}\mathcal{L}_{adv_{ds}} &= \mathbb{E}_{s\sim S}[||D_S(s) - 1||_2] + \mathbb{E}_{c\sim C}[||D_S(G_S(E_C(c)))||_2] \\ &- \lambda_{gp}\,\mathbb{E}_{\tilde{s}}[(||\nabla_{\tilde{s}} D_S(\tilde{s})||_2 - 1)^2] \\ \mathcal{L}_{adv_{dc}} &= \mathbb{E}_{c\sim C}[||D_C(c) - 1||_2] + \mathbb{E}_{s\sim S}[||D_C(G_C(E_S(s)))||_2] \\ &- \lambda_{gp}\,\mathbb{E}_{\tilde{c}}[(||\nabla_{\tilde{c}} D_C(\tilde{c})||_2 - 1)^2]\end{aligned} \tag{4.5}$$

where $\tilde{s}, \tilde{c}$ are sampled uniformly along a straight line between their corresponding domain pair of real and generated images. We set weighting factor $\lambda_{gp} = 10$.

Full learning objectives: Our full model is trained alternatively as with a standard conditional GAN framework, with the following joint optimisation:

$$\begin{aligned}&\underset{D_S, D_C}{\operatorname{argmin}}\, \lambda_{adv} L_{adv_{ds}} + \lambda_{adv} L_{adv_{dc}} \\ &\underset{E_S, E_C, G_S, G_C, D_{cls}}{\operatorname{argmin}} \quad \lambda_{embed} L_{embed} + \lambda_{recons} L_{recons} + \lambda_{adv} L_{adv_g} + \lambda_{cls} L_{cls}\end{aligned} \tag{4.6}$$

where $\lambda_{adv}, \lambda_{embed}, \lambda_{recons}, \lambda_{cls}$ are hyperparameters that control the relative importance of each loss. In this work, we set $\lambda_{adv} = 10, \lambda_{embed} = 100, \lambda_{recons} = 100$ and $\lambda_{cls} = 1$ to keep the losses in roughly the same value range.

4.3 Discriminative Factorisation for FG-SBIR

The sketch style transfer model in Sec. 4.2.1 addresses the first level of inverse-sketching by translating a sketch into a geometrically realistic contour. Specifically, for a given sketch s, we can synthesise its distortion-free sketch contour s_c as $G_C(E_S(s))$. However, the model is not

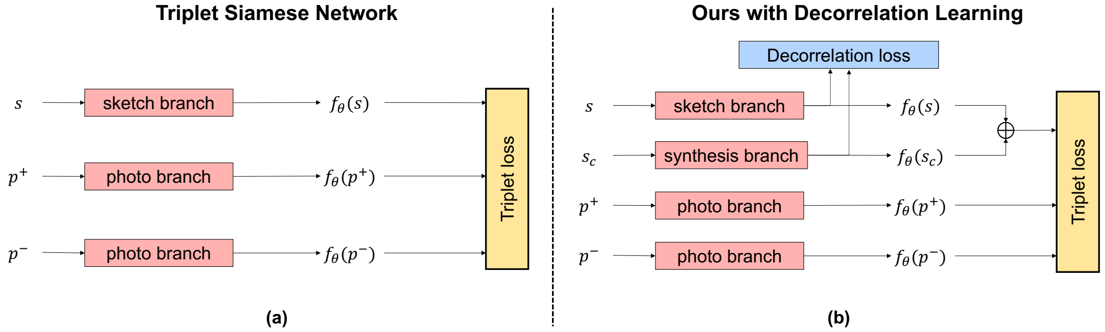


Figure 4.4: (a) Existing three-branch Siamese Network (Sangkloy et al., 2016; Yu et al., 2016) vs. (b) Our four-branch network with decorrelation loss.

trained to synthesise the sketch details inside the contour – this is harder because sketch details exhibit more subjective abstraction yet less distortion. In this section, we show that for learning a discriminative FG-SBIR model, such a partial factorisation is enough: we can take s and s_c and extract complementary detail features from s to complete the inversion process.

Problem definition: For a given query sketch s and a set of N candidate photos $\{p_i\}_{i=1}^N \in P$, FG-SBIR aims to find a specific photo containing the same instance as the query sketch. This can be solved by learning a joint sketch-photo embedding using a CNN f_θ (Sangkloy et al., 2016; Yu et al., 2016). In this space, the visual similarity between a sketch s and a photo p can be measured simply as $D(s,p) = ||f_\theta(s) - f_\theta(p)||_2^2$.

Enforcing factorisation via de-correlation loss: In our approach, clean and accurate contour features are already provided in s_c via our style transfer network defined previously. Now we aim to extract detail-related features from s. To this end we introduce a decorrelation loss to minimise the cross-covariance between $f_\theta(s)$ and $f_\theta(s_c)$:

$$L_{decorr} = ||\overline{f_\theta(s)}^T \times \overline{f_\theta(s_c)}||_F^2, \tag{4.7}$$

where $\overline{f_\theta(s)}$ and $\overline{f_\theta(s_c)}$ are obtained by normalising $f_\theta(s)$ and $f_\theta(s_c)$ with zero-mean and unit-variance respectively in a batch, and $||.||_F^2$ is the squared Frobenius norm. This ensures that $f_\theta(s)$ encodes detail-related features in order to meet the decorrelation constraint with complementary contour encoding $f_\theta(s_c)$.

Model design: Existing deep FG-SBIR models (Pang et al., 2017; Yu et al., 2016) adopt a three-branch Siamese network architecture, shown in Figure 4.4(a). Given an anchor sketch s and a positive photo p^+ containing the same object instance and a negative photo p^-, the outputs of the three branches are subject to a triplet ranking loss to align the sketch and photo in the discriminative joint embedding space learned by f_θ. To exploit our contour and detail representation, we use a four-branch Siamese network with inputs s, s_c, p^+, p^- respectively (Figure 4.4(b)). The extracted features from s and s_c are then fused before being compared with those extracted from p^+ and p^-. The fusion is denoted as $f_\theta(s) \oplus f_\theta(s_c)$, where $\oplus$ is the element-wise addition[1]. The triplet ranking loss is then formulated as:

$$L_{tri} = \max(0, \Delta + D(f_\theta(s) \oplus f_\theta(s_c), f_\theta(p^+)) - D(f_\theta(s) \oplus f_\theta(s_c), f_\theta(p^-))) \tag{4.8}$$

where Δ is a hyperparameter representing the margin between the query-to-positive and query-to-negative distances. Our final objective for discriminatively training FG-SBIR becomes:

$$\min_\theta \sum\nolimits_{t \in T} L_{tri} + \lambda_{decorr} L_{decorr} \tag{4.9}$$

we set $\Delta = 0.1, \lambda_{decorr} = 1$ in our experiments so two losses have equal weights.

4.4 Experiments

4.4.1 Experimental Settings

Dataset and preprocessing: We use the public QMUL-Shoe-V2 (Yu et al., 2017b) dataset, the largest single-category paired sketch-photo dataset to date, to train and evaluate both our sketch style transfer model and FG-SBIR model. It contains 6648 sketches and 2000 photos. We follow its standard train/test split with 5982 and 1800 sketch-photo pairs respectively. Each shoe photo is annotated with 37 part-based semantic attributes. We remove four decoration-related ones ('frontal', 'lateral', 'others' and 'no decoration'), which are contour-irrelevant and keep the rest.

[1] Other fusion strategies like element-wise multiplication have been tried and found to be inferior.

Since our style transfer model is unsupervised and does not require paired training examples, we use a large shoe photo dataset UT-Zap50K dataset (Yu and Grauman, 2014) as the target photo domain. This consists of 50,025 shoe photos which are disjoint with the QMUL-Shoe-V2 dataset. For training the style transfer model, we scale and centre the sketches and photo contours to 64×64 size, while for FG-SBIR model, the inputs of all four branches are resized to 256×256. In both experiments, to reduce the unnecessary data bias between sketch and contours, we use (Simo-Serra et al., 2018) as a mean of post-processing.

Photo contour extraction: We obtain the contour c from a photo p as follows: (i) extracting edge probability map e using (Zitnick and Dollar, 2014) followed by non-max suppression; (ii) e is binarised by keeping the edge pixels with values smaller than x, where x is dynamically determined so that when e contains many non-zero edge pixel detections, x should be small to eliminate the noisy ones, e.g., texture. This is achieved by formulating $x = e_{sort}(l_{sort} \times \min(\alpha e^{-\beta/r}, 0.9))$, where e_{sort} is the edge pixels detected in e sorted in the ascending order, l_{sort} is the length of e_{sort}, and r is the ratio between detected and total pixels. We set $\alpha = 0.08, \beta = 0.12$ in our experiments[2].. Examples of photos and their extracted contours can be seen in the last two columns of Figure 4.5.

Implementation details: We implement both models in Tensorflow with a single NVIDIA 1080Ti GPU. For the **style transfer task**: as illustrated in Figure 4.3, we denote $k * k$ conv as a $k \times k$ Convolution-BatchNorm-ReLU layer with stride 1 and $k * k$ residual as a residual block that contains two $k * k$ conv blocks with reflection padding to reduce artifacts. Upscale operation is performed with bilinear up-sampling. We do not use BatchNorm and replace ReLU activation with Tanh for the last output layer. Our discriminator has the same architecture as in (Isola et al., 2017), but with BatchNorm replaced with LayerNorm (Ba et al., 2016) since the gradient penalty is introduced. The number of discriminator iterations per generator update is set as 1. We trained for $50k$ iterations with a batch size of 64. Like **FG-SBIR task**: we fine-tune ImageNet-pretrained ResNet-50 (He et al., 2016) to obtain f_θ with the final classification layer removed. Same with (Yu et al., 2016), we enforce l_2 normalisation on f_θ for stable triplet

[2] To make it intuitive, we show values of $\min(\alpha e^{-\beta/r}, 0.9)$ here with respect to r = 0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7. They are 0.9, 0.9, 0.8819, 0.4442, 0.2656, 0.1193, 0.1017, 0.0950.

learning. We train for $60k$ iterations with a triplet batch size of 16. For both tasks, the Adam (Kingma and Ba, 2015) optimiser is used, where we set $\beta_1 = 0.5$ and $\beta_2 = 0.9$ with an initial learning rate of 0.0001 respectively.

Competitors: For style transfer, four competitors are compared. **Pix2pix** (Isola et al., 2017) is a supervised image-to-image translation model. It assumes that visual connections can be directly established between sketch and contour pairs with l_1 translation loss and adversarial training. Note that we can only use the QMUL-Shoe-V2 train split for training Pix2pix, rather than UT-Zap50K, since sketch-photo pairs are required. **UNIT** (Liu et al., 2017c) is the latest variant of the popular unsupervised CycleGAN (Kim et al., 2017; Yi et al., 2017; Zhu et al., 2017). Similar to our model, it also has a shared embedding construction subnet. Unlike our model, there is no attribute prediction regularisation and visual consistency instead of embedding consistency is enforced. **UNIT-vgg**: For fair comparison, we substitute the learned-from-scratch encoder in UNIT to our fixed VGG-encoder, and introduce the same self-residual architecture in the decoder. **Ours-attr**: This is a variant of our model without the attribute prediction task for embedding regularisation. For FG-SBIR, competitors include: **Sketchy** (Sangkloy et al., 2016) is a three-branch Heterogeneous triplet network. For fair comparison, the same ResNet50 is used as the base network. **Vanilla-triplet** (Yu et al., 2016) differs from Sketchy in that a Siamese architecture is adopted. It is vanilla as the model is trained without any synthetic augmentation. **DA-triplet** (Song et al., 2018) is the state-of-the-art model, which uses synthetic sketches from photos as a means of data augmentation to pretrain the Vanilla-triplet network and fine-tune it with real human sketches. **Ours-decorr** is a variant of our model, obtained by discarding the decorrelation loss.

4.4.2 Results on Style Transfer

4.4.2.1 Qualitative Results

Figure 4.5 shows example synthesised sketches using the various models. It shows clearly that our method is able to invert the sketching process by effectively factorising out any details inside the object contour and restyling the remaining contour parts with smooth strokes and more real-

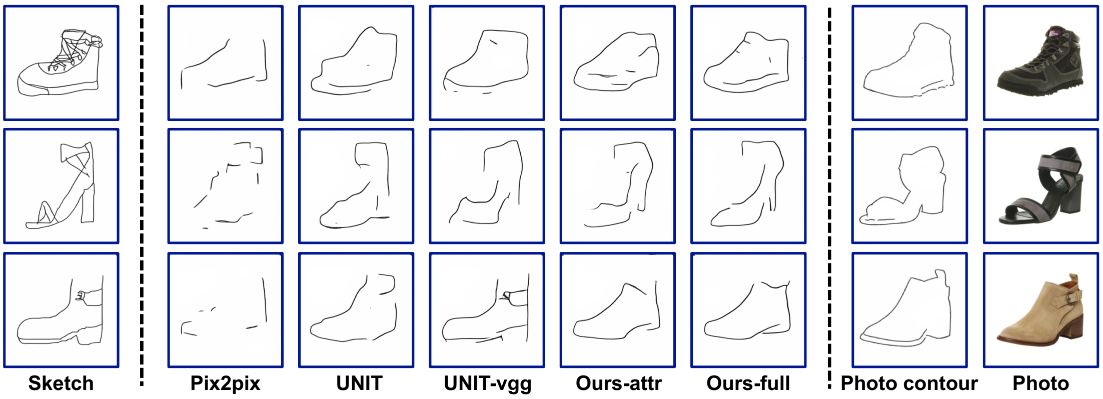


Figure 4.5: Different competitors for translating (inverting) sketching abstraction at contour-level. Illustrations shown here have never been seen by its corresponding model during training.

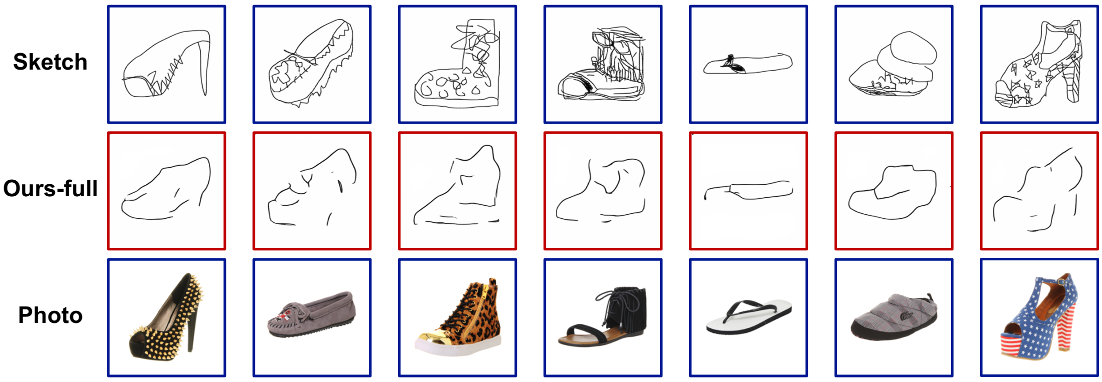


Figure 4.6: Typical failure of our model when sketching style is too abstract or complex.

istic perspective geometry. In contrast, the supervised model Pix2pix failed completely due to sparse training data and the assumption of pixel-to-pixel alignment across the two domains. The unsupervised UNIT model is able to remove the details, but struggles to keep its original identity with salient visual traits preserved, e.g., how the heel part is radically changed before and after stylisation. Using a fixed VGG-16 as encoder (UNIT-vgg) alleviates the problem but still fails to emulate the style for the object photo contours featured with smooth and continuous strokes. These results suggest that the visual cycle consistency constraint used in UNIT is too strong a constraint on the embedding subnet, leaving it with little freedom to perform both the detail removal and contour restyling tasks. As an ablation, we compare Ours-attr with Ours-full and observe that the attribute prediction task does provide a useful regularisation to the embedding subnet to make the synthesised contour more smooth and less fragmented. Our model is far from

	Chance	**Pix2pix**	**UNIT**	**UNIT-vgg**	**Ours-attr**	**Ours-full**
acc@1	0.50%	3.60%	4.50%	4.95%	6.46%	**8.26%**
acc@5	2.50%	10.51%	15.02%	17.87%	22.22%	**23.27%**
acc@10	5.00%	17.87%	26.28%	29.88%	31.38%	**35.14%**

Table 4-A: Comparative results of using the synthetic sketches obtained via different models to retrieve photos from a well-trained FG-SBIR model.

(w_c, w_n)	**UNIT vs. Ours-full**	**UNIT-vgg vs. Ours-full**	**Ours-attr vs. Ours-full**
(0.9, 0.1)	88.0%	72.0%	62.0%
(0.8, 0.2)	88.0%	70.0%	64.0%
(0.7, 0.3)	88.0%	70.0%	64.0%
(0.6, 0.4)	86.0%	68.0%	62.0%
(0.5, 0.5)	84.0%	70.0%	64.0%

Table 4-B: Pairwise comparison results of human perceptual study. Each cell lists the percentage where our full model is preferred over the other method. Chance is at 50%.

being perfect. Figure 4.6 shows some failure cases. Most failure cases are caused by the sketcher unsuccessfully attempting to depict objects with rich texture by an overcomplicated sketch. This suggests that our model is mostly focused on the shape cues contained in sketches and confused by the sudden presence of large amounts of texture cues.

4.4.2.2 Quantitative Results

Quantitative evaluation of image synthesis models remains an open problem. Consequently, most studies either run human perceptual studies or explore computational metrics attempting to predict human perceptual similarity judgements (Heusel et al., 2017; Salimans et al., 2016). We perform both quantitative evaluations. **Computational evaluation:** In this evaluation, we seek a metric based on the insight that if the synthesised sketches are realistic and free of distortion, they should be useful for retrieving photos containing the same objects, despite the fact that the details inside the contours may have been removed. We thus retrain the FG-SBIR model of (Yu et al., 2016) on the QMUL-Shoe-V2 training split and used the sketches synthesised using different style transfer models to retrieve photos in QMUL-Shoe-V2 test split. The results in Table 4-A show that our full model outperforms all competitors. The performance gap over the chance suggests that despite lack of detail, our synthetic sketches still capture instance-discriminative visual cues. The superior results to the competitors indicate the usefulness of cyclic embedding consistency and attribute prediction regularisation. **Human perceptual study:** We further evaluate our model via a human subjective study. We recruit N ($N = 10$) workers and ask each of them to perform the same pairwise A/B test based on the 50 randomly-selected sketches from QMUL-Shoe-V2 test split. Specifically, each worker undertakes two trials, where three images are given at once, i.e., a sketch and two restyled version of the sketch using two compared models. The worker is then asked to choose one synthesised sketch based on two criteria: (i) correspondence (measured as r_c): which image keeps more key visual traits of the original sketches, i.e., more instance-level identifiable; (ii) naturalness (measured as r_n): which image looks more like a contour extracted from a shoe photo. The left-right order and the image order are randomised to ensure unbiased comparisons. We denote each of the $2N$ ratings for each synthetic sketch under one comparative test as c_i and n_i respectively, and compute the correspondence measure $r_c = \sum_{i=1}^{N} c_i$, and naturalness measure $r_n = \sum_{i=1}^{N} n_i$. We then average them to obtain one score based on a weighting: $r_{avr} = \frac{1}{N}(w_c r_c + w_n r_n)$. Intuitively, w_c should be greater than w_n because ultimately we care more about how the synthesised sketches help FG-SBIR. In Table 4-B, we list in each cell the percentage of trials where our full model is preferred over the other competitors. Under different weighting combinations, the superiority of our design is consistent ($> 50\%$), drawing the same conclusion as our computational evaluation. In particular, compared with prior state-of-the-art, UNIT, our full model is preferred by humans nearly 90% of the time.

4.4.3 Results on FG-SBIR

4.4.3.1 Quantitative Results

In Table 4-C, we compare the proposed FG-SBIR model (Ours-full) with three state-of-the-art alternatives (Sketchy, Vanilla-triplet and DA-triplet) and a variant of our model (Ours-decorr). The following observations can be made: (i) Compared with the three existing models, our full model yields 14.27%, 2.41% and 2.11% acc@1 improvements respectively. Given that the three competitors have exactly the same base network in each network branch, and the same model complexity as our model, this demonstrates the effectiveness of our complementary detail representation from contour-detail factorisation. (ii) Without the decorrelation loss, Ours-decorr

Sketchy	**Vanilla-triplet**	**DA-triplet**	**Ours-decorr**	**Ours-full**
21.62%	33.48%	33.78%	33.93%	**35.89%**

Table 4-C: Comparative FG-SBIR results on QMUL-Shoe-V2 test split (Yu et al., 2017b). Retrieval accuracy at rank 1 (acc@1).

produces similar accuracy as the two baselines and is clearly inferior to Ours-full. This is not surprising – without forcing the original sketch (s) branch to extract something different from the sketch contour (s_c) branch (i.e., details), the fused features will be dominated by the s branch as s contains much richer information. The four-branch model thus degenerates to a three-branch model.

4.4.3.2 Factorisation Visualisation

We carry out model visualisation to demonstrate that $f_\theta(s)$ and $f_\theta(s_c)$ indeed capture different and complementary features that are useful for FG-SBIR, and give some insights on why such a factorisation helps. To this end, we use Grad-Cam (Selvaraju et al., 2017) to highlight where in the image the discriminative features are extracted using our model. Specifically, the two non-zero dimensions of $f_\theta(s) \oplus f_\theta(s_c)$ that contribute the most similarity for the retrieval are selected and their gradients are propagated back along the s and s_c branches as well as the photo branch to locate the support regions. The top half of Figure 4.7 shows clearly that (i) the top discriminative features are often a mixture of contour and detail as suggested by the highlighted regions on the photo images; and (ii) the corresponding regions are accurately located in s and s_c; importantly the contour features activate mostly in s_c and detail features in s. This validates that factorisation indeed takes place. In contrast, the bottom half of Figure 4.7 shows that using the vanilla-triplet model without the factorisation, the model appears to be overly focused on the details, ignoring the fact that the contour part also contains useful information for matching object instances. This leads to failure cases (red box) and explains the inferior performance of vanilla-triplet.

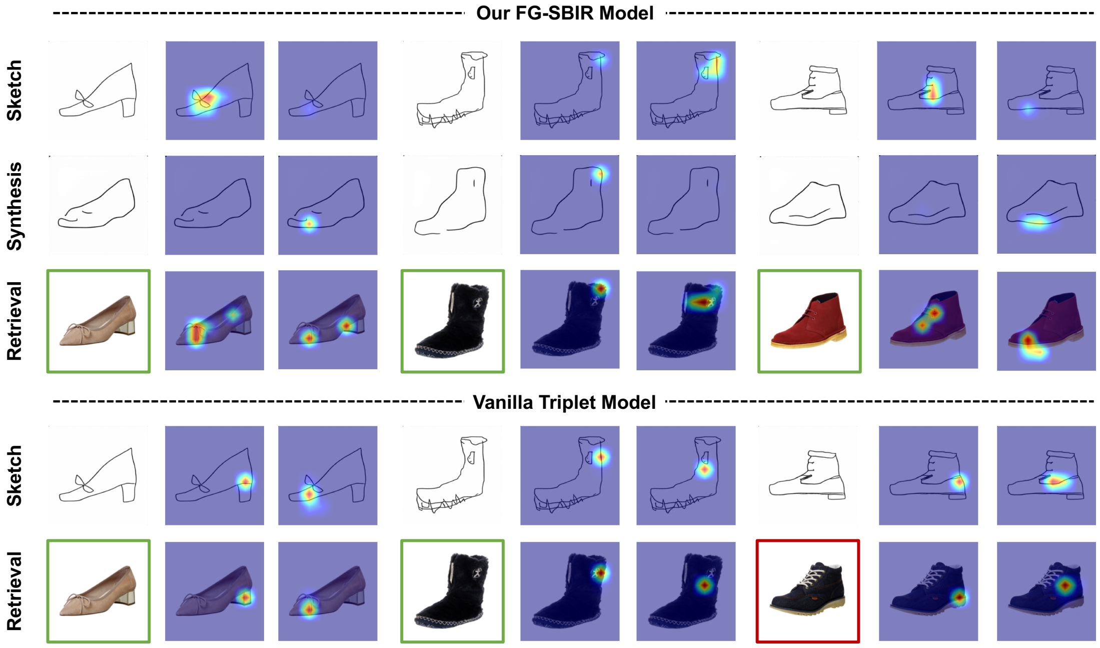


Figure 4.7: We highlight supporting regions for the top 2 most discriminative feature dimensions of two compared models. Green and red borders on the photos indicate correct and incorrect retrieval, respectively.

4.5 Summary

This chapter for the first time has defined and identified the problem of factorised inverse-sketching as a key for both sketch modelling and sketch-photo matching. Given a sketch, our deep style transfer model learns to factorise out the details inside the object contour and invert the remaining contours to match more geometrically realistic contours extracted from photos. We subsequently develop a sketch-photo joint embedding which completes the inversion process by extracting distinct complementary detail features for FG-SBIR. We demonstrated empirically that our style transfer model is more effective compared to existing models thanks to a novel cyclic embedding consistency constraint. We also achieve state-of-the-art FG-SBIR results by exploiting our sketch inversion and factorisation.

Chapter 5

Generalising FG-SBIR

5.1 Background and Motivation

In this chapter, we aim to improve cross-category generalisation in FG-SBIR. Existing work have thus far implicitly assumed that instance-level annotations of positive and negative pairs are available for every coarse category to be evaluated. This assumption limits the practical applicability of FG-SBIR. More specifically, as we shall show in this paper, in practice FG-SBIR generalises very poorly if training and testing categories are disjoint. This is of course unsatisfactory for potential users of FG-SBIR such as e-commerce, where it would be desirable to train a FG-SBIR system once on an initial set of product categories, and then have it deployed directly to newly added product categories – without needing to collect and annotate new data and retrain the FG-SBIR model. Compared to other category-level tasks such as object recognition in photo images, this annotation barrier is particularly high for FG-SBIR as instance-specific sketches are expensive and slow to collect.

To understand why the existing FG-SBIR models have limited cross-category generalisation ability, consider that the task of FG-SBIR as essentially binary classification – to differentiate corresponding and non-corresponding sketch-photo tuples. In this sense, a change of *category* is a domain-shift (Csurka, 2017) from the perspective of the machine learning model trained to perform matching. For example, a model trained on fine-grained matching of car photos and sketches, would struggle to perform fine-grained matching of bicycle images, due to inexperience with handlebars and saddles. Exposed to such out-of-sample data, the triplet-trained sketch/photo embedding networks may no longer place matching images nearby and vice-versa. Having identified the challenge as one as domain-shift, this suggests two categories of approaches to alleviating this issue: (i) Unsupervised domain *adaptation* approaches (Csurka, 2017; Ganin et al., 2016) would use unlabelled target data to adapt the model to better suit the target data; and (ii) domain *generalisation* approaches (Shankar et al., 2018) aim to train a model that is robust enough to immediately generalise to the new domain's data off-the-shelf. We address the harder domain generalisation setting – due to the practical value of not requiring target domain (category) data collection and model retraining.

We propose a new framework that automatically adapts the deep feature extraction to a given query sketch. This ensures a good representation is produced at testing-time, even when dealing with out-of-sample data in the form of sketches and photos from novel categories. The key idea is to learn an auxiliary unsupervised embedding network that maps any given sketch to a universal dictionary of prototypical sketch traits or manifold embeddings. We call this universal because it is a representation that cuts across categories. This network can thus be used to provide a latent visual trait descriptor (VTD) of any sketch (from either a training or novel category). This descriptor is in turn used to paramaterise both photo and sketch feature extractors to adapt them to the current query sketch category. Figure 5.1 illustrates the unsupervised embedding learned by our auxiliary network via an illustrative five (of 300) learned embeddings (dictionary words). One can see how categories (such as flowers) span multiple embeddings and how individual embeddings group thematically similar sketches. For example descriptor 2 and 140 encompass "complicated-dense" and "simple-sparse" visual patterns for flowers and trees; while descriptor 207 and 249 model "leftwards full-body view" and "frontal face view" respectively for cows and horses. We can also see how both training (left subgroups) and disjoint testing sketch category

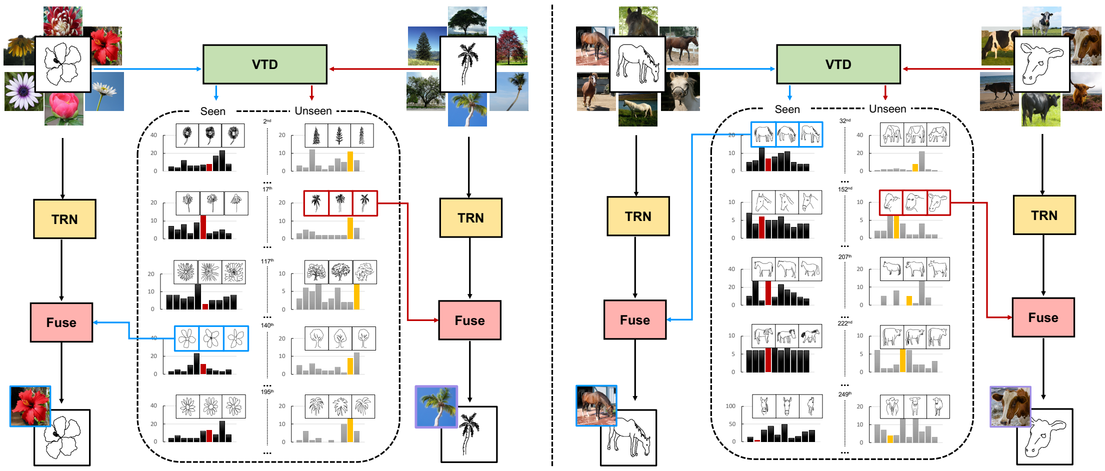


Figure 5.1: Illustration of our proposed method using four categories, organised into two related pairs. TRN: triplet ranking network. VTD: visual trait descriptor. In each bar-type VTD, we visualise its ten top distributed categories and highlight the specific one along with three representative sketch exemplars. Each sketch is uniquely assigned to one VTD that describes a category-agnostic abstract sketch trait, which is in turn used to dynamically paramaterise the TRN so as to adapt it to the query sketch. See how both training and testing sketches thematically and coherently mapped to some shared VTDs. Best viewed in colour and zoom, more details in text.

(right subgroups) are assigned to the same descriptor according to common sketch traits.

The introduction of this auxiliary universal embedding network is inspired by the pioneering *Noise As Targets* (NAT) (Bojanowski and Joulin, 2017) model. NAT proposes to pre-generate the set of all embeddings randomly – as noise – and then learn a network to map the data to this fixed noise distribution. However NAT approximately solves a cumbersome and costly discrete assignment problem to match images with embeddings at each back-propagation iteration. In contrast, we propose a novel approach to learning an embedding network based on the Gumbel-Softmax (Jang et al., 2017) reparameterisation trick. As a result, the learning is faster and more stable; and more flexible in that several alternative objectives can be considered in the same formulation. Overall our framework can be considered as a solution to domain generalisation (Shankar et al., 2018) that adapts a model via a domain-descriptor, but where the descriptor is estimated from a single data instance rather than assuming it is given as metadata (Yang and Hospedales, 2015, 2016); and where the perspective on descriptor definition is one of latent-domain discovery (Xu et al., 2014).

5.2 Methodology

5.2.1 Overview

Our framework consists of two main components. Firstly, our *unsupervised embedding network* maps any sketch s into one of K unique visual trait descriptors D_s via an encoder-decoder framework $D_s = \phi(s)$. So the full set of M-dimensional trait descriptors defines a matrix $D \in \mathbb{R}^{K \times M}$. This serves to provide the description of any sketch's query domain. Secondly, a *dynamically paramaterised feature extractor with triplet loss* is formulated, which actually performs FG-SBIR by using the generated descriptor to adapt the feature extraction and retrieval to any query sketch. Denoting $\psi(\cdot)$ as Deep CNN feature extractor, FG-SBIR is performed by finding the photo p that minimises the distance $d_{\phi(s)}(s, p) = ||\psi_{\phi(s)}(s) - \psi_{\phi(s)}(p)||_2^2$ to query sketch s. The unsupervised embedding network is trained in an unsupervised way on the training sketch categories. And the dynamically paramaterised FG-SBIR model is trained in a supervised way on the training sketch categories. No components touch the held out testing category data until evaluation. In the following sections we describe each component in detail.

5.2.2 Universal Visual Trait Embedding

5.2.2.1 Embedding for Categorical Variables

The unsupervised embedding network will map any sketch to an entry in a dictionary of descriptors D. Inspired by NAT (Bojanowski and Joulin, 2017), we pre-generate the descriptor dictionary at random so that each row of D, denoted D_i is sampled from the standard Gaussian and then ℓ_2 normalised. This ensures that the descriptor dictionary spans the available M dimensional space well. The network's goal is then to learn to map any sketch onto one of these K (random) dictionary elements so that the representations of the full sketch dataset spread out over the whole embedding space.

Encoder-Decoder We start by feeding an input sketch s into a CNN encoder $E(s)$. We then use one fully-connected (FC) layer to predict a K-dimensional vector of unnormalised probabilities p and select the most probable one as sketch s's descriptor D_s out of the full dictionary

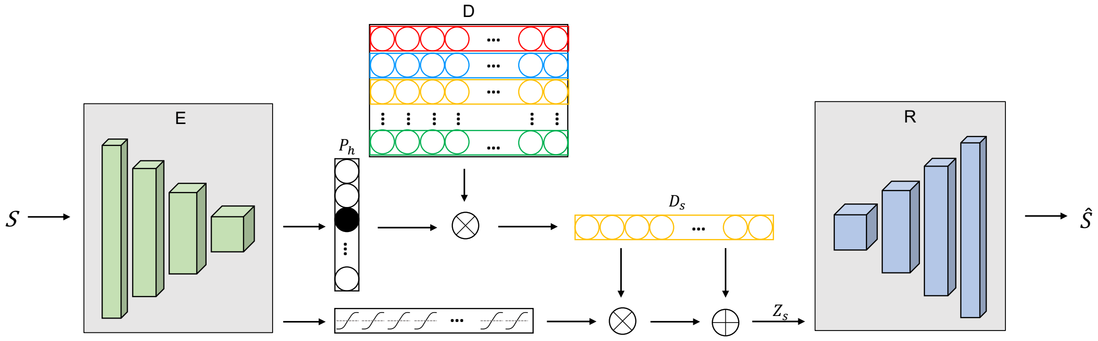


Figure 5.2: Schematic illustration of our proposed unsupervised encoder-decoder model. D is a dictionary of descriptors that represent universal sketch visual traits. Since self-reconstruction is the only learning signal here, we introduce skip-connection that generates an instance-specific perturbation of the chosen dictionary element to help ease optimisation, and in turn benefits dictionary learning.

D:

$$\begin{aligned} & p = W_p E(s) + b_p \\ & p_h = \text{onehot}(\text{argmax}(\text{softmax}(p))) \\ & D_s = p_h D, \quad \hat{s} = R(D_s) \end{aligned} \tag{5.1}$$

To ensure that each descriptor corresponds to a visually meaningful trait, the assigned descriptor is then decoded by decoder R with de-convolutional layers that reconstruct the input sketch $\hat{s} \approx s$. We denote the extraction of a sketch trait descriptor in this way as $D_s = \phi(s)$.

A Practical Consideration Since the number of descriptors K (300) is much less than sketches (tens of thousands), our approach means that sketches will be coarsely quantised, and reconstruction error will be high. (The clusters do not contain enough information to accurately reconstruct each sketch). Therefore we modify this approach with the following skip connection to improve the decoding via R.

$$\begin{aligned} & Z_s = D_s(1 + \alpha tanh(W_{sk} E(s) + b_{sk})) \\ & \hat{s} = R(Z_s) \end{aligned} \tag{5.2}$$

where we set $\alpha = 0.02$. This passes through some detailed features of the sketch to augment the coarse dictionary encoding. See Figure 5.2 for an intuitive illustration.

5.2.2.2 Optimisation for argmax

The method as presented so far is hard to optimise because: (i) The use of argmax is non-differentiable and would naively require Monte Carlo estimates and a REINFORCE-type algorithm (Williams, 1992), which suffers from high variance. (ii) A trivial minimiser of the reconstruction loss is to output one or few constant one-hot vectors p_h. Especially in the early phase of training, this will trap the model in a local minima forever. To alleviate this problem, we employ a low-variance gradient estimated based on a reparameterisation trick.

Hard Assignment via Gumbel-Softmax Applying the Gumbel-Softmax reparameterisation trick (Jang et al., 2017) and straight-through (ST) gradient estimator, p_h is replaced as:

$$\begin{aligned} p_g &= \operatorname{softmax}((p+g)/\tau) \\ p_{hg} &= \operatorname{onehot}(\operatorname{argmax}(p_g)) \end{aligned} \tag{5.3}$$

where $g \in \mathbb{R}^K$ with $g_1 ... g_k$ are i.i.d samples drawn from $\operatorname{Gumbel}(0,1)$, and τ is the temperature[1]. We further enforce a uniform categorical prior on $p_s = \operatorname{softmax}(p)$ to avoid sketches being assigned to only a subset of dictionary elements, and form a Kullback-Leibler loss as:

$$\begin{aligned} q_y &= [1/K, 1/K, ..., 1/K] \in \mathbb{R}^K \\ D_{\mathrm{KL}}(p_s || q_y) &= \frac{1}{B} \sum_{i=1}^{B} \mathbf{p_s}_{i,:} \log(\mathbf{p_s}_{i,:}/q_y) \end{aligned} \tag{5.4}$$

where B is the batch size. For simplicity, we use bold $\mathbf{p_s}$ to denote the batch counterpart of p_s, with $\mathbf{p_s}_i$ the i^{th} example and $\mathbf{p_s}_{i,j}$ as its j^{th} element. We will follow this convention for other symbols. This ensures that across the batch as a whole, sketches are encouraged to assign to diverse descriptors.

Soft Assignment via Entropy Constraint We also explore an alternative strategy, which is to adopt a soft assignment approach during training. By replacing p_h with p_s, each sketch takes

[1] For the forward pass, p_{hg} is used thus a real one-hot vector is generated, while for the backward pass, it is replaced by p_g to make the (estimated) gradient flows back. In practice, we just assign it a mild value like 1.0 instead of an annealing strategy as in (Jang et al., 2017).

a linear combination of D, rather than selecting a row of D for representation learning. In this soft assignment of sketches to descriptors, we want to motivate sparse probabilities so that each s tends to receive one dominant label assignment. Thus we add a row entropy loss:

$$\mathrm{H}_{\text{row}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{K}\mathbf{p_s}_{i,j}\log(\mathbf{p_s}_{i,j}) \tag{5.5}$$

Eq. 5.5 achieves its minimum 0 only if $\mathbf{p_s}_i$ is an one-hot vector specifying a deterministic distribution. We further encourage equal usage of all $\mathbf{p_s}_{:,j}$ via a column entropy term:

$$\begin{aligned} p_c &= \frac{1}{B}\sum_{i=1}^{B}\mathbf{p_s}_{:,j} \in R^K \\ \mathrm{H}_{\text{col}} &= -\sum_{j=1}^{K} p_{c_j} log(p_{c_j}) \end{aligned} \tag{5.6}$$

Eq. 5.6 achieves its maximum 1 only if elements in p_c are uniformly distributed. However, the row entropy constraint is only valid for a large enough minibatch and we empirically find that on average around 30% of p_h are still empty, (no assignments of any sketches). Therefore, we dynamically replace the stale and inactive D_i during training and bring them back in to compete with over-active ones. Specifically, we extract p_h of all training sketches after each epoch, and select the most concentrated D_i. A small random perturbation is then added to define a new centre, i.e., $D_i(1 + \beta\mathcal{N}(0, 1))$. We find this simple strategy works well[2].

5.2.2.3 Full Objectives

Depending on which assignment strategy we use (Gumbel-Softmax vs. Entropy), and combined with reconstruction loss $L_{\text{rec}} = ||s - \hat{s}||_2$, we obtain our two optimisation objectives:

$$\begin{aligned} &\min \mathbb{E}_{s\sim S}[L_{\text{rec}} + \lambda_{\text{KL}} D_{\text{KL}}(p_s||q_y)] \\ &\min \mathbb{E}_{s\sim S}[L_{\text{rec}} + \lambda_{\text{row}}\mathrm{H}_{\text{row}} - \lambda_{\text{col}}\mathrm{H}_{\text{col}}] \end{aligned} \tag{5.7}$$

[2] A side effect is to trade quality with time. We spend almost one-third of the time extracting representations for all training sketches. We set $\beta = 0.05$ throughout the experiments and find it works well empirically.

where hyper-parameters $\lambda_{\text{KL}}, \lambda_{\text{row}}, \lambda_{\text{col}}$ control the relative weighting importance. In summary, optimising the unsupervised objective Eq. 5.7 trains an autoencoder that internally represents sketches in terms of a pre-defined K-element dictionary D. In the following section, we will re-use the sub-network that assigns sketches to dictionary elements $D_s = \phi(s)$ as a descriptor for dynamically paramaterising our FG-SBIR network.

5.2.3 Dynamic Parameterisation for FG-SBIR

The unsupervised embedding network shown in Figure 5.2 extracts a visual trait descriptor (VTD), $\phi(s)$, from each sketch, which is then used to parameterise a triplet ranking network (TRN), $\psi(\cdot)$, for learning domain-generalisable representations for sketch and photo, as illustrated in Figure 5.1. Note that sketch and photo feature extractors ψ is Siamese – applied to both sketch and photo for FG-SBIR. Denoting $\psi_{\phi(s)}(\cdot)$ as the feature extractor calibrated to sketch s, and $F(\cdot)$ as a vanilla CNN feature extractor, we have:

$$\psi_{\phi(s)}(\cdot) = \eta(\phi(s)) \odot F(\cdot) + F(\cdot) \tag{5.8}$$

The above can be interpreted as a small hypernetwork (Ha et al., 2017), where we generate a sketch-conditional diagonal weight layer to adapt the conventional CNN feature F to the current sketch, along with a residual connection. It can also be interpreted as a generating a sketch-specific soft attention mask on F where η indicates salient dimensions. Using this dynamically paramaterised feature extractor, we finally apply a standard triplet loss to match photos and sketches:

$$\begin{aligned} L_{\text{tri}} = \max(0, \Delta + d(\psi_{\phi(s)}(s), \psi_{\phi(s)}(p^+)) \\ - d(\psi_{\phi(s)}(s), \psi_{\phi(s)}(p^-))) \end{aligned} \tag{5.9}$$

A Stochastic Paramaterisation A standard solution for the weight generator $\eta(\cdot)$ in Eq. 5.8 is to transform the input sketch embedding through a few FC layers (Ha et al., 2017). However, as the input is a discrete set of descriptor vectors, this causes discontinuity in weight generation. We take inspiration from (Zhang et al., 2017) and mitigate this by introducing layers that predict a Gaussian mean and variance, and then sample these to more smoothly generate the target parameters. This yields more training pairs to encourage robustness to small perturbations along the conditioning manifold.

$$\begin{aligned}\mu_s &= W_\mu \phi(s) + b_\mu \\ \sigma_s &= \exp(\frac{W_\sigma \phi(s) + b_\sigma}{2}) \\ \eta(\phi(s)) &= \mu_s + \sigma_s \odot \mathcal{N}(0,1).\end{aligned} \tag{5.10}$$

Optimisation and Inference Finally, to further enforce the smoothness over the conditioning manifold and avoid overfitting (Doersch, 2016), we add the commonly applied variational regularisation term, $L_{\text{con}} = D_{\text{KL}}(\eta(\phi(s))||\mathcal{N}(0,I))$, weighted by a small value λ_{con}. Our FG-SBIR objective is:

$$\min \mathbb{E}_{t \sim T}[L_{\text{tri}} + \lambda_{\text{con}} L_{\text{con}}] \tag{5.11}$$

where t stands for a triplet tuple, consisting of $\{s, p^+, p^-\}$. During testing, for a query sketch s, we sample $\eta(\phi(s))$ ten times to calculate distance for each sketch-photo gallery pair and take the smallest as the final measure.

5.3 Experiments

5.3.1 Experimental Settings

Dataset and Pre-processing We use the public Sketchy (Sangkloy et al., 2016) and QMUL-Shoe-V2 (Yu et al., 2017b) to evaluate our methods. Sketchy contains 125 categories with 100 photos each and at least 5 sketches per photo. We follow the same dataset split as (Yelamarthi et al., 2018) and partition Sketchy into 104 train and 21 test categories to ensure the test ones are not present in 1000 ImageNet Challenge classes (Russakovsky et al., 2015). For QMUL-Shoe-V2, we test generalisation by transferring between fine-grained sub-categories and design five groups of such experiments as shown in Table 5-C. We scale and centre the sketches to 64×64 when training VTD, while for FG-SBIR, the inputs of all three branches are resized to 299×299.

Implementation Details We implement both models in Tensorflow on a single NVIDIA 1080Ti

GPU. For unsupervised embedding network: our CNN-based encoder-decoder, E and R, contains five stride-2 convolutions and five fractional-convolutions with stride 1/2, with one 1×1 convolution at the end and start of each. BatchNorm-Relu activation is applied to every convolutional layer, except the output of R with Tanh. All hyper-parameters are set to undergo a warm-up phase, so that reconstruction loss dominates the training at the beginning. We train the models for 200 epochs under all settings with λ_{kl}, λ_{row}, λ_{col} linearly increasing from $0, 1, 1$ to $1.5, 2, 10$ respectively. The dictionary D has $M = 256$ dimensions and $K = 300$ elements throughout. We use Adam optimiser with learning rate 0.0002. For FG-SBIR: we fine-tune ImageNet-pretrained Inception-v3 (Szegedy et al., 2016) to obtain F with the final classification layer removed. We enforce ℓ_2 normalisation on the output of η to stabilise triplet learning and set hyper-parameters $\Delta = 0.1, \lambda_{con} = 0.004$. We train for 20 epochs on Sketchy, and 10 epochs on QMUL-Shoe-V2 with a learning rate of 0.0001 and Adam optimiser under all settings.

Evaluation Metric We use Acc.@ K to measure the FG-SBIR performance, which is the percentage of sketches whose true-match photos are ranked in the top K.

5.3.2 Competitors

Sketchy If not otherwise mentioned, all competitors are implemented based on Inception-v3, and our model is trained with soft assignment. **Hard-Transfer** (Yu et al., 2016) trains a vanilla Siamese triplet ranking model and is directly tested on unseen categories. **CVAE-Regress**[3] (Yelamarthi et al., 2018) is the state-of-the-art zero-shot SBIR method by learning a conditional generative model to regress ImageNet-pretrained photo features to their corresponding sketch features. **Reptile** (Alex and Johnn, 2018) is a recent meta-learning algorithm that repeatedly samples tasks, trains them, and moves the initialisation towards the trained weights. We integrate it in (Yu et al., 2016) by each time randomly sampling 52 categories to form two subtasks and train parallelly for 500 iterations. **CrossGrad** (Shankar et al., 2018) is a state-of-the-art domain generalisation method that trains both a label and a domain classifier on examples perturbed by each other's loss gradients. For our task, we regard each of 104 training cate-

[3] This method is designed for category-level characterisation, so is expected to perform poorly. The reason we didn't adapt it to embrace triplet ranking loss is that even with it, since the photo features are fixed rather than learned, some poor performance is still naturally expected.

Competitor	**Acc.@ 1**	**Acc.@ 5**	**Acc.@ 10**	**Competitor**	**Acc.@ 1**	**Acc.@ 5**	**Acc.@ 10**
Hard-Transfer	16.0%	40.5%	55.2%	**Ours-WordVector**	18.0%	43.5%	58.7%
CVAE-Regress	2.4%	9.5%	17.7%	**Ours-Classify**	16.2%	41.4%	57.2%
Reptile	17.5%	42.3%	57.4%	**Ours-Full/Edge**	16.8%	41.3%	56.2%
CrossGrad	13.4%	34.9%	49.4%	**Ours-Full/Hard**	20.1%	46.4%	61.7%
Ours-VAE	12.7%	34.5%	49.7%	**Ours-Full**	**22.6%**	**49.0%**	**63.3%**
Ours-VAE-Kmeans	17.6%	41.9%	56.9%	**Upper-Bound**	29.9%	65.5%	81.4%

Table 5-A: Comparative Cross-Category FG-SBIR results on Sketchy (Sangkloy et al., 2016).

gories as a unique domain and 100 inter-category photo ids as labels. **Ours-VAE** corresponds to training a conventional variational autoencoder (VAE) (Kingma and Welling, 2013) without our visual trait descriptor and using the per-instance latent representation as the descriptor ϕ to parameterise the FG-SBIR model. **Ours-VAE-Kmeans** performs K-means clustering in the VAE latent space, to generate a dictionary of sketch descriptors analogous to our approach, but without end-to-end learning. **Ours-WordVector** and **Ours-Classify** replace our descriptor with the category-level semantics driven descriptor either drawn from the class name (Mikolov et al., 2013) or extracted from the penultimate feature layer of a sketch classification network. Lastly, we compare our proposed model (**Ours-Full**) with its two ablated versions, including **Ours-Full/Hard** and **Ours-Full/Edge**, which are trained with hard assignment strategy instead of soft, on edgemaps other than human freehand sketches respectively.

QMUL-Shoe-V2 This is a single category product-level FG-SBIR dataset. We do not have enough data to train a dictionary D from scratch. Therefore we take the advantage of the best visual trait descriptor trained on Sketchy and introduce two variants **Ours-Sketchy** and **Ours-Sketchy-Ft**. They differ in if we directly use the Sketchy dictionary or further fine-tune it on the seen sub-category of QMUL-Shoe-V2. **Hard-Transfer** is the competitor.

Caveat Since we use all images within one category for constructing a challenging test set. The **Upper-Bound** for both datasets is therefore likely a slight overestimate, as it uses half of these for training before before testing on all.

5.3.3 Results on Sketchy

5.3.3.1 Comparison with Competitors

We compare the performance of different models in Table 5-A and make the observations: (i) The gap between direct transfer (16%) and a model trained using data from the target (unseen) categories (Upper-Bound, 30%) is large, confirming the cross-category generalisation gap. (ii) Our model beats all 10 competitors in bridging this gap. (iii) For DG meta-learning competitors, CrossGrad fails to improve on the direct transfer baseline, but Reptile does improve on it. However both are worse than our full model. (iv) Comparing our two proposed optimisaton methods, soft assignment outperforms hard. We attribute this to the rigid approach of the latter – it enforces a uniform distribution over assignment to descriptors, which may not hold in practice since some will be more common than others. (v) Our visual trait descriptor approach is beneficial as manifested by the dramatic performance gap between ours and the conventional VAE, VAE-Kmeans alternatives in particular. (vi) Using visually abstract but neat human free-hand sketches as source data to train our descriptor is important. Replacing these with the detailed but noisy edgemaps extracted from natural photos hurts the performance. This suggests that the model is able to exploit the clean and iconic free-hand sketches to learn abstract visual traits more effectively.

5.3.3.2 Qualitative Impact of Descriptors

We now qualitatively examine how a visual trait descriptor $D_s = \phi(s)$ impacts sketch photo matching and how retrieval is affected if using another sketch descriptor $D_{\hat{s}}, \hat{s} \neq s$ instead. To achieve this, we select one dimension from $\psi_{\phi(s)}$ that contributes the most to successful matching and use Grad-Cam (Selvaraju et al., 2017) to propagate gradients back to highlight discriminative image regions. This can be seen as a visualisation of the implicit attention mechanisms that different visual trait descriptors define to adapt the feature extraction. We illustrate this in Figure 5.3 across five different D_ss for each of six sketch-photo pairs. It shows that (i) The corresponding D_s helps focus attention on regions with similar spatial support for both s and p^+, while a mismatched $D_{\hat{s}}$ fails to do this; (ii) Individual descriptors D_i are useful for multiple categories,

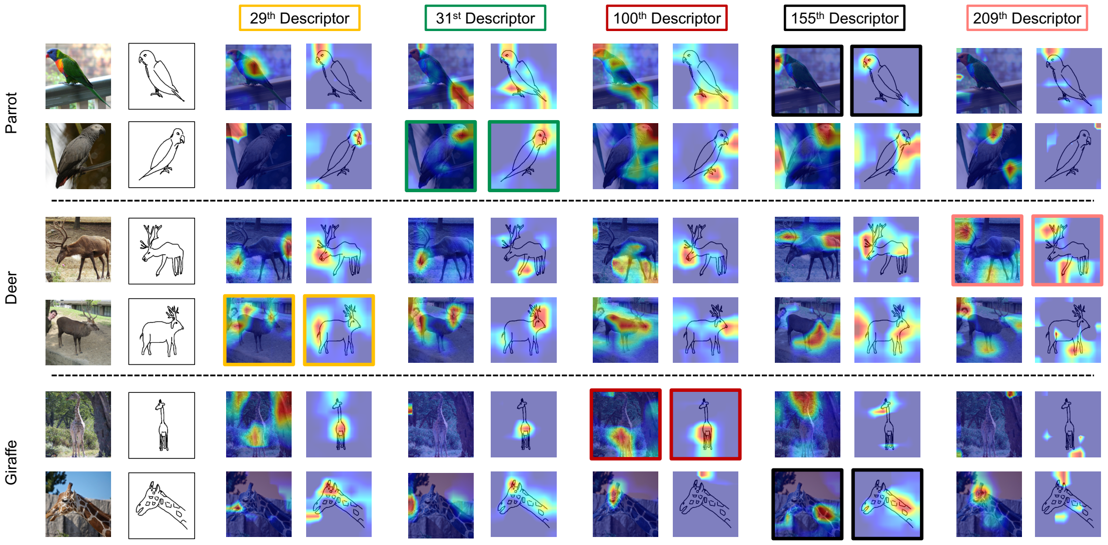


Figure 5.3: Visualisation of how the VTD adapts the sketch-photo matching process. Coloured image box border indicates when the correct (corresponding to query sketch) descriptor is used to paramaterise the embedding space.

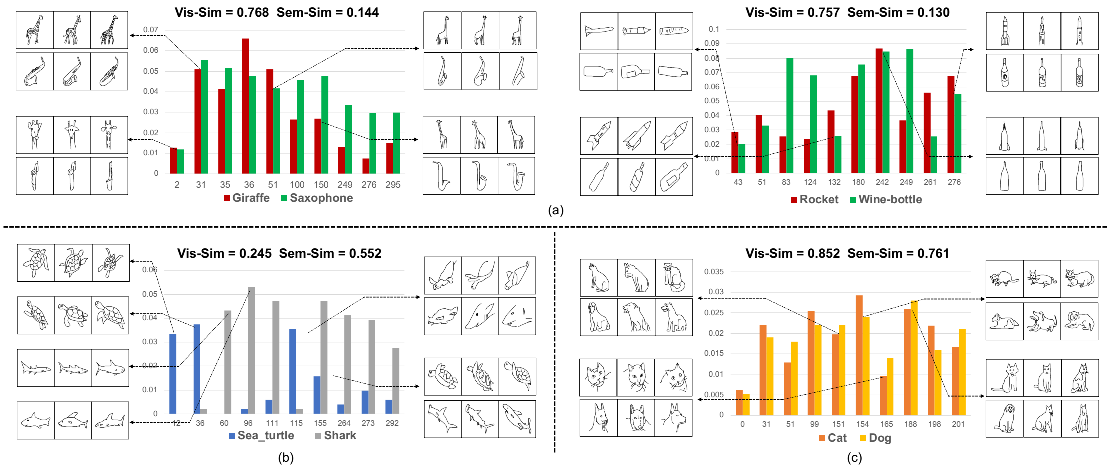


Figure 5.4: Word-Vector vs. Visual-Semantics. Comparing illustrative category pairs: (a) Visually close but semantically far.(b) Semantically related but visually far. (c) Visually and semantically related. Vis-Sim is cosine distance between the histograms, and Sem-Sim is the cosine distance between word-vectors. Histograms shown here are the ten most similar descriptors jointly shared between two categories. Best viewed in colour and zoom.

e.g., the 155^{th} descriptor for parrots and giraffes.

No.	Hard Assignment			Soft Assignment		
	Acc.@ 1	**Acc.@ 5**	**Acc.@ 10**	**Acc.@ 1**	**Acc.@ 5**	**Acc.@ 10**
20	18.4%	43.3%	58.4%	19.5%	46.0%	60.4%
100	19.6%	45.7%	60.9%	20.7%	47.7%	62.7%
300	**20.1%**	**46.4%**	**61.7%**	**22.6%**	**49.0%**	**63.3%**
1000	17.8%	42.3%	57.6%	18.3%	43.8%	59.0%

Table 5-B: Effects of different number of VTDs on Cross-Category FG-SBIR performance on Sketchy dataset(Sangkloy et al., 2016).

5.3.3.3 How Many Descriptors?

We investigate the impact of the descriptor dictionary size K on CC-FG-SBIR performance in Table 5-B. We can see that our model is not very sensitive to K under either hard and soft assignment strategies, and a few hundred suffices for good performance.

5.3.3.4 Descriptor-Category Spread

We can verify that VTDs cross-cut rather than mirror the category breakdown of sketches. On average, training sketches from each category are assigned to 138 ± 30 unique descriptors. Testing category sketches (upon which the embedding is not trained) are assigned to 129 ± 33 descriptors, indicating that the cross-cutting spread is retained despite the train/test domain-shift.

5.3.3.5 Word-Vector vs. Visual-Semantics

The quantitative results (Table 5-A) showed that word-vector descriptors do improve performance over hard-transfer, albeit much less than our approach. We can contrast similarity as estimated by word-embeddings, with that of our VTD. Figure 5.4(a) shows a pair of categories which are far in semantic word similarity, but near in visual visual trait descriptor similarity. Here *category level* visual similarity is measured by the number of sketches (y-axis) from different categories (bars) co-assigned to a single descriptor (x-axis). In contrast, Figure 5.4(b) shows semantically related categories that are visually distinct (shark/sea turtle) and Figure 5.4(c) illustrates categories that are both semantically and visually related (dog/cat).

Sub-category	**Fine-grained Transfer**	**No. Train / Test**	**Competitor**	**Acc.@ 1**	**Acc.@ 5**	**Acc.@ 10**
Sandal	Flat → Wedge	560 / 227	Hard-Transfer	9.25%	32.2%	48.0%
			Ours-Sketchy	13.2%	34.3%	50.4%
			Ours-Sketchy-Ft	**15.4%**	**37.9%**	**54.6%**
			Upper-Bound	28.6%	56.8%	72.2%
Toe-shape	Closed → Fish-mouth	400 / 351	Hard-Transfer	14.8%	44.7%	61.5%
			Ours-Sketchy	22.2%	50.4%	65.0%
			Ours-Sketchy-Ft	**24.2%**	**54.5%**	**66.7%**
			Upper-Bound	29.3%	56.7%	71.8%
Shoe-height	Ankle- → Knee-high	2010 / 245	Hard-Transfer	10.6%	32.2%	43.3%
			Ours-Sketchy	14.7%	38.0%	51.0%
			Ours-Sketchy-Ft	**18.4%**	**40.8%**	**55.1%**
			Upper-Bound	25.3%	54.3%	71.8%
Heel-shape	Thick → Thin	828 / 411	Hard-Transfer	12.2%	35.0%	48.7%
			Ours-Sketchy	15.1%	**41.4%**	**59.4%**
			Ours-Sketchy-Ft	**17.3%**	41.1%	57.7%
			Upper-Bound	26.3%	61.8%	80.5%
Topline	Small → Big	5015 / 1543	Hard-Transfer	7.25%	22.9%	34.5%
			Ours-Sketchy	12.2%	28.9%	39.7%
			Ours-Sketchy-Ft	**15.5%**	**31.4%**	**43.8%**
			Upper-Bound	19.6%	44.2%	61.5%

Table 5-C: Comparative FG-SBIR results on generalising between sub-categories on QMUL-Shoe-V2 dataset (Yu et al., 2017b)

5.3.4 Results on QMUL-Shoe-V2

In this section, we borrow the best VTD dictionary D (Ours-Full) trained on Sketchy and use it to help transfer between sub-categories in QMUL-Shoe-V2. To test generalisation on this benchmark, we design five groups of experiments, each defining a different type of train/test gap, and with diverse split sizes. We report their performance in Table 5-C and find that compared with Hard-Transfer, even when directly applying D to this novel dataset, Ours-Sketchy improves performance in all experiments. This is promising as a Sketchy-trained dictionary is generally applicable and it has potential to benefit other specific FG-SBIR applications. When further fine-tuned on the *train* data split of each experiment, we also usually improve performance (Ours-Sketchy-Ft vs. Ours-Sketchy).

5.4 Summary

This chapter for the first time identified the generalisation problem in cross-category FG-SBIR and proposed a novel solution via learning a universal visual trait descriptor embedding. This embedding dictionary is mapped to a set of latent domains that cross-cut sketch categories, and enable a retrieval network to be suitably paramaterised given a query sketch – by mapping query sketches to the corresponding descriptor in the dictionary. Extensive experiments on Sketchy and QMUL-Shoe-V2 demonstrate the superiority of our proposed method for cross-category FG-SBIR.

Chapter 6

Solving Mixed-modal Jigsaw Puzzle for FG-SBIR

6.1 Background and Motivation

In this chapter, we aim to propose a self-supervised pre-training alternative to ImageNet pre-training which is long considered to be critical for promising FG-SBIR performance. Despite the great strides made, almost all contemporary competitive FG-SBIR models depend crucially on one necessary condition: the model must be fine-tuned from the pre-trained weights of an ImageNet (Deng et al., 2009) classifier. The reason behind this is that collecting instance-level sketch-photo pairs for FG-SBIR is very expensive, with the largest current single product-category dataset being only on a scale of thousands. Scaling such data collection to the size required to train a contemporary deep CNN from scratch is infeasible. Thus, ImageNet pre-training is ubiquitously leveraged to provide initialisation for FG-SBIR.

While useful in ameliorating the otherwise fatal lack of data for FG-SBIR, ImageNet pre-training suffers from mismatch to the intended downstream task. Training for object category classification requires detecting high-level primitives that characterise different object categories, while learning to ignore certain fine-grained details critical for the instance-level recognition

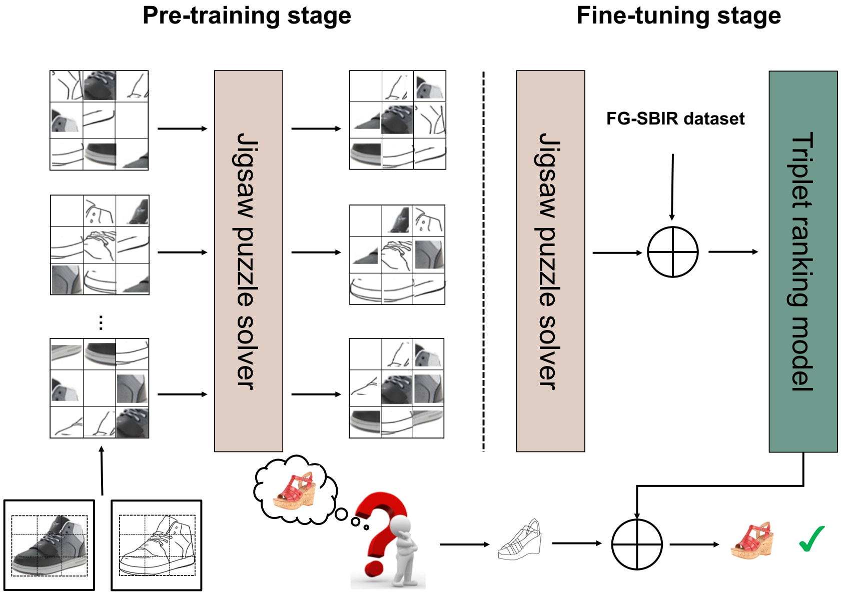


Figure 6.1: Conventionally, a competitive FG-SBIR system relies on two prerequisites: ImageNet pre-training and triplet fine-tuning. Here we investigate substituting the former with a mixed-domain jigsaw puzzle solver, leading to improved FG-SBIR accuracy and generalisation.

task in FG-SBIR. Crucially, ImageNet only contains images from the photo modality, while FG-SBIR requires cross-modality matching between photo and sketch. This suggests that ImageNet classification may not be the most effective pre-training strategy for FG-SBIR. Indeed, recently (Radenovic et al., 2018) explored the self-supervised task of matching a photo with its edgemap to substitute the sketch-photo pair for model training. This could potentially be used for pre-training as well. However, its effectiveness is limited because the task boils down to edge detection and is not challenging enough for the model to learn fine-grained cross-modal discriminative patterns for matching.

We propose to perform representation pre-training by recovering an image from mixed-modal shuffled patches. That is, patches drawn randomly from photo and edgemap domains. Solving this problem, as illustrated in Figure 6.1, requires learning to bridge the domain discrepancy, to understand holistic object configuration, and to encode fine-grained detail in order to characterise each patch accurately enough to infer their relative placement.

Note that jigsaw solving has been studied before (Carlucci et al., 2019; Noroozi and Favaro, 2016) for single-modal recognition problems. In this work, differently, we deal with a more challenging mixed-modal jigsaw problem. Solving jigsaw puzzle as a task itself is hard; as a result, instead of directly solving it, i.e., recovering the un-shuffled original image where all patches are put back to the right places, most prior work (Carlucci et al., 2019; Kim et al., 2018; Noroozi and Favaro, 2016) poses jigsaw solving as a recognition task. In contrast, we frame the jigsaw solving problem as a permutation inference problem and solve it using Sinkhorn iterations (Adams and Zemel, 2011; Santa Cruz et al., 2017). Our experiments show that this formalisation of a jigsaw solver provides a much stronger model for self-supervised representation pre-training on all four publicly available product-level FG-SBIR datasets. A surprising outcome is that this approach can completely break the category associations between representation pre-training and FG-SBIR fine-tuning without harming performance, as well as lead to improved generalisation across categories between FG-SBIR fine-tuning and run-time testing stage.

6.2 Jigsaw Pre-training for FG-SBIR

6.2.1 Overview

This section aims to introduce a self-supervised pre-training strategy in the form of solving mixed-modal jigsaw puzzles. The whole FG-SBIR training pipeline thus consists of two stages: self-supervised jigsaw pre-training and supervised FG-SBIR triplet fine-tuning. The first self-stage will use photos p and corresponding programmatically produced edgemaps e to produce mixed modal jigsaw images x. Our jigsaw solver $J(x)$ trains a representation by learning to solve these jigsaws. In the second stage, we use the learned representation as an initial condition, and fine-tune a FG-SBIR model by supervised triplet ranking on annotated pairs of free-hand sketches and photos.

6.2.2 Jigsaw Puzzle Generation

We first define a cross-modality shuffling operator $x = T(e, p, O, R)$, that transforms a photo p and its edgemap counterpart e to form a mixed-modal jigsaw image x. Assume the jigsaw image is to contain N patches in a $\sqrt{N} \times \sqrt{N}$ array. O is then a random permutation of an array $[1 \dots N]$ that describes the mapping of input image patches to the jigsaw patches in x, and R

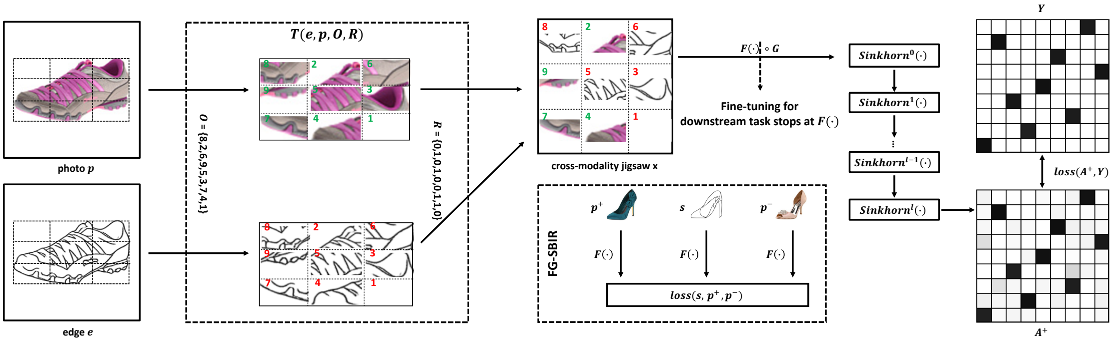


Figure 6.2: Schematic of our proposed Jigsaw pre-training for FG-SBIR. We take a jigsaw puzzle of 9 tiles as an example. Both photo p and its edgemap counterpart e are first divided into 3×3 grid and reshuffled based on a permutation order O. Using a random binary vector R, these are then stitched into the final mixed-modality jigsaw x. x is fed to our jigsaw solver $J(x) = G(F(x))$ including a ConvNet feature extractor $F(\cdot)$ and Sinkhorn-based permutation solver $G(\cdot)$ to obtain the permutation matrix A^+ that solves the jigsaw. After pre-training, we take the CNN module $F(\cdot)$ and use it as a feature extractor for FG-SBIR fine-tuning.

is a N-dimensional vector of Bernoulli samples that will determine whether input patches are drawn from photo p or edgemap e. Thus, as shown in Figure 6.2, x is generated by drawing the ith patch from location O_i of the inputs, specifically from sketch if $R_i = 1$ and photo if $R_i = 0$.

6.2.3 Jigsaw Puzzle Solver

Our jigsaw solver $J(x)$ processes the mixed-modal jigsaw image x and returns A^+, a $N \times N$ assignment matrix that maps each jigsaw patch to the target patch of an un-shuffled image (Figure 6.2). The jigsaw solver $J(x) = G(F(x))$ is implemented via a CNN feature extractor $F(\cdot)$, followed by a permutation solver $G(\cdot)$. The solver applies a fully connected layer W on the CNN's output to produce an affinity matrix $A \in \mathbb{R}^{N \times N}$, where A_{ij} describes the CNNs preference strength for assigning the i^{th} input puzzle location to the j^{th} target location. It then infers the most likely global assignment of jigsaw patches to output patches by applying the Sinkhorn operator to the affinity matrix $A^+ = \mathrm{Sinkhorn}(A)$. This will help to un-shuffle the input patches and solve the jigsaw by producing an assignment matrix with constraint[1]: (i) all elements are either 0 or 1; (ii) each row and column has exactly one assignment. For instance,

[1] Strictly speaking, the Sinkhorn iterator only guarantees a "soft" assignment matrix, i.e., a doubly stochastic matrix where each row and column adds up to 1, and reach to a strict assignment matrix by further framing it as a maximisation problem via argmax. But in practice even with such relaxation, we found that optimised by cross-entropy loss, $\mathrm{Sinkhorn}(A)$ can already approximate a strict assignment matrix well and saves the efforts to deal with non-differentiable issue with argmax.

$A^+_{ij} = 1$ means assigning i^{th} input patch to the j^{th} target patch, and the mapping between input and output patches is 1-to-1.

Sinkhorn Operator $\text{Sinkhorn}(\cdot)$ To implement the Sinkhorn operator, we follow (Adams and Zemel, 2011) and iteratively normalise its rows of the input in order to approximate the doubly stochastic matrix A^+:

$$\begin{aligned}
&\text{Sinkhorn}^{\,0}(A) = \exp(A) \\
&\text{Sinkhorn}^{\,l}(A) = T_c(T_r(\text{Sinkhorn}^{\,l-1}(A))) \\
&\text{Sinkhorn}(A) = \lim_{l \to \infty} \text{Sinkhorn}^{\,l}(A)
\end{aligned} \tag{6.1}$$

where $T_r(X) = X \oslash (X\mathbf{1}_N\mathbf{1}_N^T)$, $T_c(X) = X \oslash (\mathbf{1}_N\mathbf{1}_N^T X)$ as the row and column-wise normalisation operations of a matrix, with $\oslash$ denoting the element-wise division and $\mathbf{1}_N$ a column vector of ones. l is a hyper-parameter to control the number of Sinkhorn iterations used to estimate the assignment.

Loss Functions For jigsaw pre-training, our loss function aims to close the distribution gap between A^+ and the true assignment matrix Y (generated from from O), defined as:

$$\text{loss}(A^+, Y) = -\sum_{i=1}^{N}\sum_{j=1}^{N}[\log(A^+_{ij}) \times Y_{ij} + \log(1 - A^+_{ij}) \times (1 - Y_{ij})] \tag{6.2}$$

Summary At each iteration, training images are edge extracted, and randomly shuffled and modality mixed. Training the jigsaw solver J to un-shuffle the images requires the CNN to learn a feature extractor which is both modality invariant, and encodes enough fine-grained detail to enable the permutation solver to successfully un-shuffle.

6.3 FG-SBIR Fine-Tuning

In the fine-tuning stage we perform supervised learning of free-hand sketch to photo retrieval. Specifically, we strip off the permutation solver module G and use the feature extractor $F(\cdot)$ in the standard triplet ranking loss:

$$\begin{aligned}&\text{loss}(s, p^+, p^-) = \\ &\max(0, \Delta + d(F(s), F(p^+)) - d(F(s), F(p^-)))\end{aligned} \tag{6.3}$$

where s is a query sketch, p^+ and p^- are positive and negative photo examples, $d(s,p) = ||F(s) - F(p)||_2^2$, and Δ is a hyper-parameter as the margin between the positive and negative example distance. For evaluation we retrieve the photo p with minimum distance to a query sketch s according to $d(s,p)$.

6.4 Experimental Settings

To pinpoint the advantages of jigsaw pre-training, we control all baselines and ablated variants to use the same CNN architecture and optimisation strategy. Learning rates and hyper-parameters are not grid-searched for optimal performance. Only training iterations may vary across datasets.

Dataset and Pre-processing **For Jigsaw pre-training**: The FG-SBIR benchmarks used are the Shoe, Chair and Handbag product search datasets from (Yu et al., 2017b). For pre-training, additional photo images of the same category are collected. (1) Shoes – we take all 50,025 product images from (Yu and Grauman, 2014). (2) Handbags – we randomly select 50k photos from Handbag-137k (Zhu et al., 2016) which is crudely crawled from Amazon without manual refinement. We filter out the ones with noisy background or irrelevant visuals, e.g., a handbag with a human model, which leaves a final size of 42,734. (3) Chair – we collect chair images from various sources to assure their diversity, including MADE, IKEA and ARGOS, and contribute 7,813 chair photos overall. We take 90% of these photos for self-supervised training, and use the rest as validation for model selection. We extract edgemaps from photos using (Zitnick and Dollar, 2014). **For Triplet fine-tuning**: We use all four publicly available product FG-SBIR datasets (Yu et al., 2017b) to evaluate our methods, namely QMUL_Shoe_V1, QMUL_Shoe_V2, QMUL_Chair and QMUL_Handbag, with 419, 6,648, 297, 568 sketch-photo pairs respectively. Of these, we use 304, 5,982, 200, 400 pairs for training and the rest for testing following the same splits as in (Yu et al., 2017b). Since noticeable data bias exists between edgemaps for pre-training and sketches in fine-tuning, e.g., stroke width, blurriness, we process both sketches and edgemaps via a cleanup and simplification model (Simo-Serra et al., 2018). We scale and centre all input images at both stages on a 256x256 blank canvas before feeding into a model.

Implementation Details All experiments are carried out with a base architecture $F(\cdot)$ of GoogleNet (Szegedy et al., 2015) running on Tensorflow with a single NVIDIA 1080Ti GPU. **For Jigsaw pre-training**: the initial learning rate is set to 1e-3 for 50k iterations and decreased to 1e-4 for another 10k with a batch size of 128. Since product images have white background, it's likely when dividing it into a $N \times N$ grid that some corner patches will be completely empty. Thus in practice, we first draw bounding boxes around the object (by simple pixel-value thresholding) in both photo and edgemap images and perform patch shuffling within them. The number of iterations l for the Sinkhorn operator is set to $5, 10, 15, 20$ for the patch number $N = 4, 9, 16, 25$ respectively. Intuitively, denser jigsaws pose more complicated un-shuffling problems and thus require more Sinkhorn iterations. To discourage overfitting to patch-edge statistics (Noroozi and Favaro, 2016), we leave a random gap between the patches. For **Triplet fine-tuning**: We train triplet ranking with a batch size of 16. We train 50k iterations for QMUL_Shoe_V2 and 20k iterations for the rest. The learning rate is set 1e-3 with a fixed margin value $\Delta = 0.1$. As a run-time augmentation, we also adopt the multi-cropping strategy as in (Yu et al., 2016). In both stages, common training augmentation approaches including horizontal flipping and random cropping, as well as colour jittering are applied. `MomentumOptimizer` is used with momentum value 0.9 throughout.

Evaluation Metrics Following community's convention, FG-SBIR performance is quantified by acc@K, the percentage of sketches whose true-match photos are ranked in top K. We focus on the most challenging scenario of K=1 through our experiments. Each experiment is run five times. The mean and standard deviation of the results obtained over the five trials are then reported.

Baselines As our focus is on pre-training, our baselines consist of alternative pre-training approaches, while the final triplet fine-tuning is kept the same throughout. **Counting** (Noroozi et al., 2017) and **Rotation** (Gidaris et al., 2018): These are two popular self-supervised alternatives to Jigsaws. The former asks for the total number of visual primitives in each split tile to equate that in the whole image. The latter requires the model to recognise the 2d rotation applied to an image. We found the common 2x2 split for learning to count may seemingly suffice for categorisation purpose, but empirically too coarse for fine-grained matching. Therefore in our implementation, we enhance it to count within 3x3 split, which is equivalent to training a 11-way Siamese network (9 tiles + 1 original image + 1 contrastive negative image[2] to circumvent trivial learning). We follow the same definition of geometric rotation set (Gidaris et al., 2018) by multiples of 90 degrees, i.e., 0, 90, 180, and 270 degrees, which makes a 4-way classification objective. **Contrastive Predictive Coding (CPC)** (Oord et al., 2018): A state of the art self-supervised method that predicts the representations of patches below a certain position from those above it via autoregressive model. This is learned by correctly classifying the "future" representation amongst a set of unrelated negative representations. We follow the authors' implementations by predicting up to five rows from the 7×7 grid. **Matching**: This trains a triplet ranking model between an edgemap query and the positive and negative photo counterparts (Radenovic et al., 2018). **ImageNet (Szegedy et al., 2015)**: this corresponds to the standard pre-trained 1K classification model on ImageNet, GoogleNet in our case. **Our/1000-way**: we adapt our mixed-modality jiasaw solving based model, but instead of solving it, we follow (Kim et al., 2018; Noroozi and Favaro, 2016) to solve a substitute problem of 1000-way jigsaw pattern classification. Lastly, **Ours** and **Ours/ImageNet**, two means of training our proposed method either from scratch or building upon the initialised weights of ImageNet.

6.5 Results and Analysis

6.5.1 Comparison with Baselines

Our first discovery is that self-supervised jigsaw pre-training from scratch on *target* category photos (i.e., For FG-SBIR on shoe products, collect un-annotated shoe photos for pre-training) followed by standard FG-SBIR fine-tuning is highly effective. Belows is more detailed analysis

[2] A potential shortcut is that it can easily satisfy the constraint by learning to count as few visual primitives as possible, so many entries of the feature embedding may collapse to zero without a contrastive signal.

Pre-training		FG-SBIR Dataset			
Method	**Self-supervised?**	**QMUL_Shoe_V1**$^{4\times4}$	**QMUL_Shoe_V2**$^{3\times3}$	**QMUL_Chair**$^{3\times3}$	**QMUL_Handbag**$^{4\times4}$
Counting	✓	41.74%± 2.30	30.42%± 0.54	72.78%± 4.35	54.05%± 2.77
Rotation	✓	32.17%± 2.68	28.83%± 0.40	70.31%± 3.45	38.33%± 1.86
CPC	✓	21.91%± 1.69	8.65%± 0.34	35.24%± 0.42	15.36%± 0.69
Matching	✓	39.13%± 0.87	31.05%± 0.84	75.69%± 1.53	50.36%± 0.68
ImageNet	✗	43.48%± 1.74	33.99%± 1.09	85.16%± 1.56	52.62%± 2.04
Ours/1000-way	✓	42.78%± 3.75	30.24%± 1.74	79.59%± 1.53	49.40%± 3.97
Ours/ImageNet	✗✓	48.00%± 2.91	31.26%± 0.65	79.59%± 1.34	61.07%± 1.50
Ours	✓	**56.52%± 2.75**	**36.52%± 0.84**	**85.98%± 2.01**	**62.97%± 2.04**

Table 6-A: Comparisons with different baselines as pre-training approaches for FG-SBIR task. The top-right superscript on each dataset name indicates the granularity of the jigsaw game solved that brings the best FG-SBIR performance respectively.

of the results with reference to Table 6-A.

Is solving a cross-modality jigsaw task a better strategy than ImageNet pre-training? Yes. It is evident that the proposed method (Ours) outperforms all the other baselines including the conventional ImageNet pre-training based one (ImageNet) on all four datasets, sometimes with significant margins. Furthermore ImageNet pre-training does not provide any benefits, but harmful when combined with our jigsaw solver (Ours/ImageNet). These results show that training for single-modality object classification is of limited relevance compared to our mixed-modal pre-training strategy.

Does the way the jigsaw puzzle is solved matter? Yes. The significant gap between Ours and Ours/1000-way confirms the significance of our technical choice: Formalising jigsaw solving as permutation estimation via Sinkhorn operator to actually solve it. This difference in efficacy is due to two reasons: (i) How to choose the pre-defined permutation set for classification determines the ambiguity of the task. Despite efforts to maximise task efficacy via evolution of classification sets (Noroozi and Favaro, 2016), classifying among a fixed set of permutations is worse than our assignment matrix estimation which must select among *all possible* permutations. (ii) The Sinkhorn operator provides a direct representation and estimation of permutation, so that latent features are properly learned to support this purpose, rather than a coarse correlate to permutation.

Why is Edge-photo-matching ineffective? At the first glance, training an edge-photo matching model (Radenovic et al., 2018) seems a natural task choice for pre-training FG-SBIR, given the similarity between edges and human sketches[3]. However, the very poor performance of the baseline (Matching) suggests that even though the edgemap is useful substitute to sketch (as demonstrated by our method), how to design the cross-modal task matters. The Edge-photo-matching task only requires whole image level photo to edgemap matching, which can be effectively solved by learning an edge detector. In contrast, our mixed-modal jigsaw puzzle problem is much harder – solving it requires the model to understand the two modalities both at the image level and the local patch level.

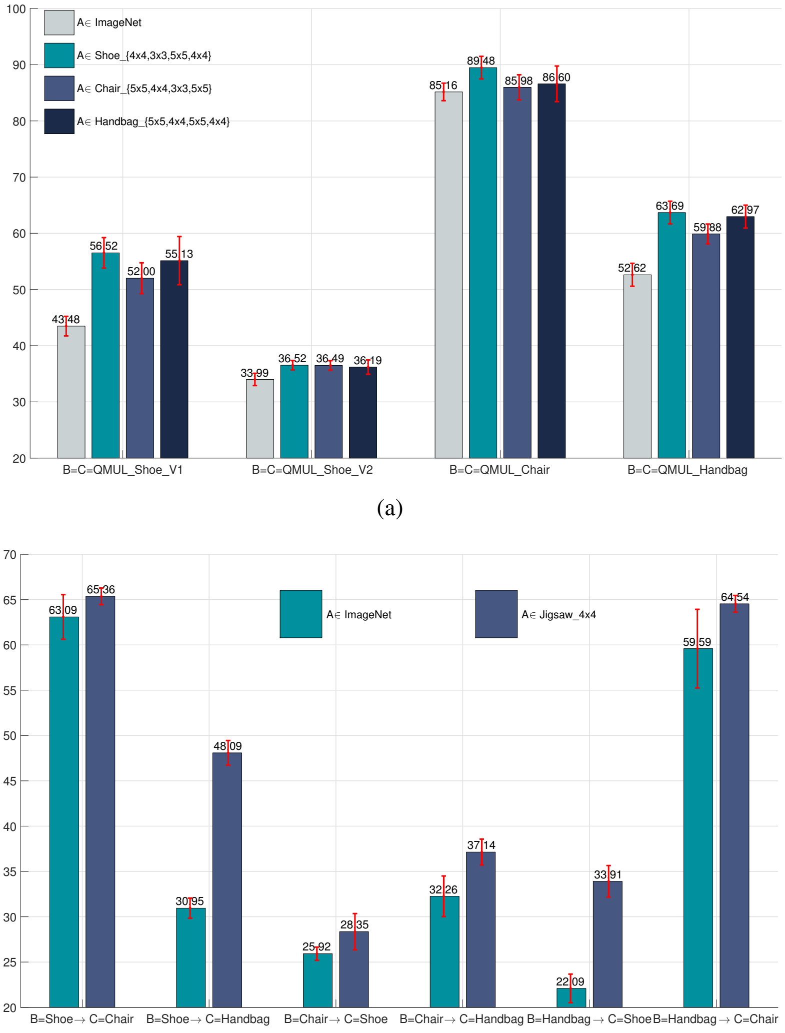
Why do the improvements vary across datasets? It is noted that our method exhibits a bigger margin on the shoes and handbags compared to chairs. Although our pre-training task is well aligned with the downstream SBIR-task, data sourced from different categories is likely to shape the model's behaviour in different ways. We believe overall solving jigsaw puzzles on shoes and handbags are harder than chairs due to the more complicated and diverse design styles they present, and thus better model capabilities are required and gained through the jigsaw solving pre-training stage.

6.5.2 Cross-Category Generalisation of Jigsaws

Our second discovery is that models pre-trained to solve jigsaw puzzles are surprisingly generalisable. Pre-training on one category followed by triplet fine-tuning and testing on another category is similar or sometimes even better compared with two stages within the same category.

Analysis of Jigsaw-informed Pre-training Model We first investigate the importance of having the same object category during jigsaw pre-training and triplet fine-tuning stages. From the results in Figure 6.3(a), we make the following observations: (i) Matching pre-training and fine-tuning category is not crucial. Indeed using the Shoe dataset for pre-training tends to provide the best performance across all four fine-tuning/testing categories. (ii) This suggests what really matters is not whether the pre-train/fine-tune categories are aligned, but the richness of each individual pre-training dataset itself. In this regard we observe Shoe>Handbag>Chair in

[3] Indeed, especially in the field of image-to-image translation, people tend to treat the terms sketch and edgemap interchangeably.



(a)

(b)

Figure 6.3: Cross-Category generalisation in pre-training and FG-SBIR. Symbols A, B, C refer to FG-SBIR model learning pattern A+B→C, where A represents our jigsaw training data, further fine-tuned by a triplet ranking model on category B, and finally testing on category C. We slightly abuse the notation here, as sometimes A can also be ImageNet. We use the notation $=$ to denote using the same category for two of these stages. (a) Cross-category generalisation between jigsaw pre-training and fine-tuning/testing. Fine-tuning/testing is kept the same throughout (B=C). (b) Cross-category generalisation between pre-training/fine-tuning and testing. Pre-training/fine-tuning are kept the same throughout (A=B). Best viewed in zoom.

terms of which dataset provides the most effective pre-training across a variety of target datasets. This result also coincides with our intuition that a good pre-training model should be category-agnostic. (iii) Overall, as long as pre-training uses our proposed jigsaw strategy, and is provided with a moderate sized set of product photos from any fashion category, the standard ImageNet pre-training strategy can be beaten. A key implication of these results are to provide a new route to scaling FG-SBIR systems in practice. While collecting large annotated free-hand sketch-photo pair datasets for each object category is prohibitively expensive, collecting product photos in any fashion category at large scale is quite feasible and can be used to boost FG-SBIR performance.

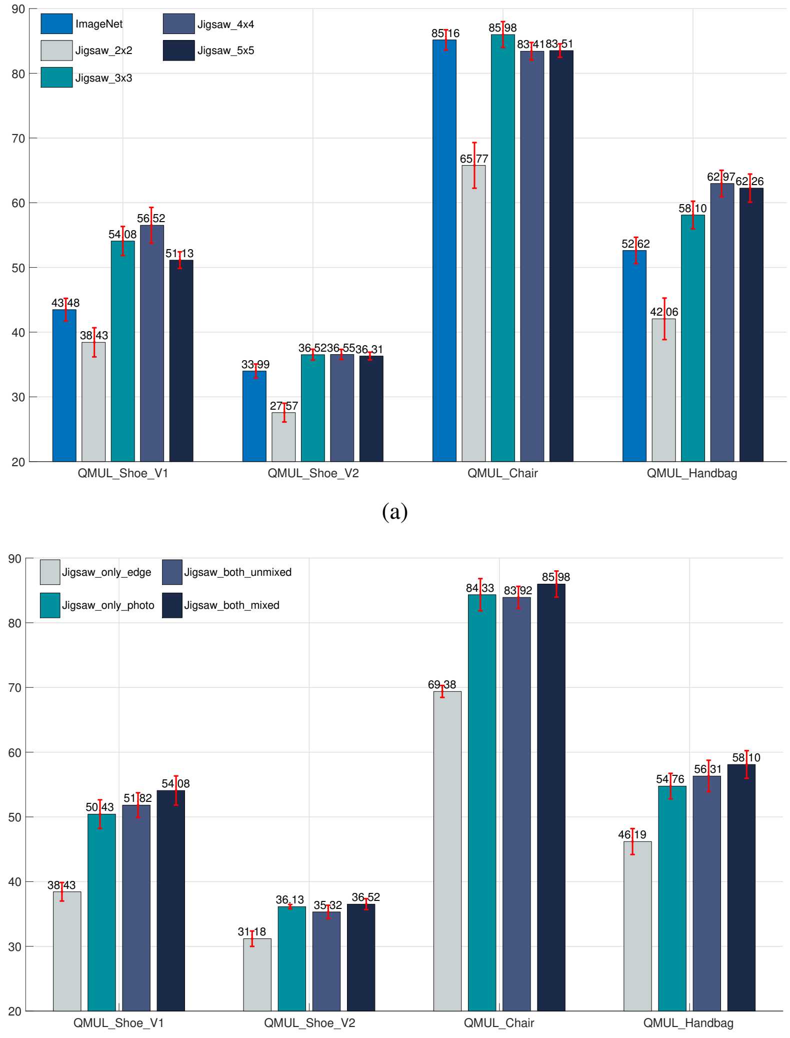
Analysis of Jigsaw-enabled FG-SBIR Model A second type of generalisability we explore is the impact of the chosen pre-training approach on the ability of the resulting FG-SBIR model to transfer across categories between training and testing. From the results in Figure 6.3(b), we can see that as expected, the performance drops in this cross-category testing setting compared to Figure 6.3(a). However, in every case Jigsaw pre-training leads to better cross-category generalisation than standard ImageNet pre-training.

6.5.3 Ablation Study

In this section, we compare our proposed method with a few variants to validate some key design choices in our jigsaw puzzles pre-training paradigm.

Granularity of Puzzle The difficulty of the jigsaw game depends on the granularity of the pieces shuffled for recomposition. If the granularity is very coarse, e.g., 2×2, the task is relatively simple and may not pose sufficient challenge for effective feature learning. If the granularity is very fine, e.g., 10×10, it may be too hard for even humans to solve and lead to models overfitting on noise. We explore this effect by enumerating jigsaw sizes from 2×2 to 5×5 and show the results in Figure 6.4(a). We make the following observations: (i) Except for 2×2, the difference in FG-SBIR results across different granularities is small and all larger jigsaws usually outperform the ImageNet baseline. (ii) The optimal granularity of jigsaw pre-training for each dataset slightly differs, but generally a puzzle of 3×3 or 4×4 provides a good choice.

Construction of Puzzle Modality Given the collected photos and extracted edgemaps of one category, there are four ways to construct the modality of the pre-training puzzles, namely: photo domain only, edgemap domain only, photo and edgemap mixed at image-level (both modalities of images are used, but each puzzle only contains a single randomly chosen modality), photo



(a)

(b)

Figure 6.4: Comparisons between different ablated variants of the proposed jigsaw pre-training on the performance of FG-SBIR task – (a) Granularity of the jigsaw. (b) Data modality of the image. The red error bar represents the standard deviation among the five repeated trials. More details in text. Best viewed in zoom.

and edgemap mixed at patch-level (ours). We summarise the results of these variants in Figure 6.4(b) and draw some conclusions: (i) Although our downstream task is cross-domain, pre-training on photo domain only seemingly sufficient for quite good performance across datasets. This is in contrast to using edgemaps alone where performance plummets. (ii) Mixing photo and edgemap images into a single dataset of both modalities provides limited benefit over photo only (Jigsaw_both_unmixed). (iii) Our patch-wise mixed-modal input strategy (Jigsaw_both_mixed) leads to the best performance on all four datasets.

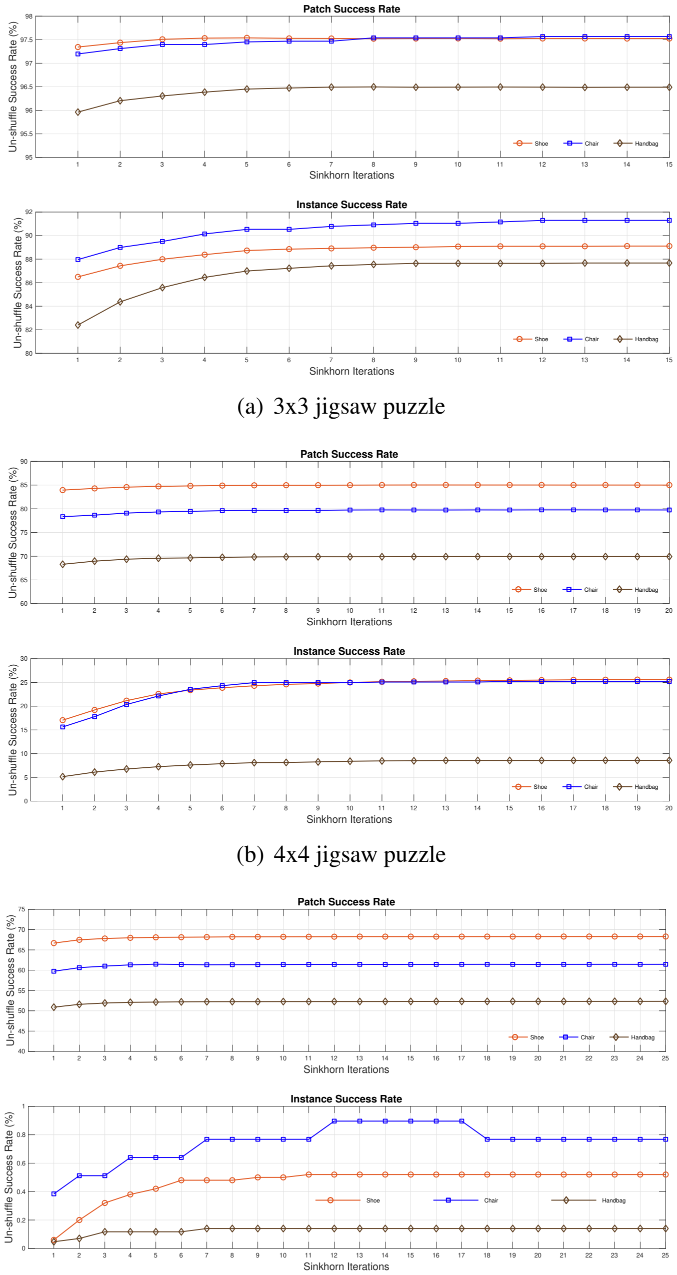
Datasets	**Variants**	**Methods**	**Acc@1**
QMUL_Shoe_V1	**C2FF**	ImageNet	44.57%± 1.58
		Ours$^{shoe\_4\times4}$	**55.30%±2.27**
	HOLEF	ImageNet	44.18%± 2.25
		Ours$^{shoe\_4\times4}$	**54.61%± 1.13**
	UFG-SBIR	ImageNet	26.96%± 1.74
		Ours$^{shoe\_4\times4}$	**35.30%± 2.92**
QMUL_Chair	**C2FF**	ImageNet	83.30%± 1.85
		Ours$^{shoe\_4\times4}$	**91.54%± 1.98**
	HOLEF	ImageNet	85.77%± 2.24
		Ours$^{shoe\_4\times4}$	**89.90%± 1.34**
	UFG-SBIR	ImageNet	**72.37%± 2.35**
		Ours$^{shoe\_4\times4}$	**72.16%± 2.53**
QMUL_Handbag	**C2FF**	ImageNet	**57.14%± 2.59**
		Ours$^{shoe\_4\times4}$	**57.38%± 2.21**
	HOLEF	ImageNet	54.29%± 1.70
		Ours$^{shoe\_4\times4}$	**63.33%± 2.68**
	UFG-SBIR	ImageNet	32.86%± 2.03
		Ours$^{shoe\_4\times4}$	**56.43 %± 0.98**

Table 6-B: Comparisons between our jigsaw approach and ImageNet pre-training when using different FG-SBIR variants.

6.6 Further Discussions

Sensitivity to Existing FG-SBIR Frameworks Thus far we have focused entirely on different pre-training approaches and datasets, while keeping a standard CNN and FG-SBIR matching architecture to facilitate direct comparison. We next examine to what extent our pre-training methods complement recent improvements in FG-SBIR method design. We consider three FG-SBIR variants, including: (i) Architecture enhancements: Coarse to fine fusion (Song et al., 2017b; Yu et al., 2017a), which we denote `C2FF`; (ii) Training objective: (Song et al., 2017b): Triplet ranking loss with a higher order learnable energy function - `HOLEF`; (iii) Problem formulation: Unsupervised FG-SBIR - `UFG-SBIR`, where edgemap is treated as a human sketch for SBIR training (Radenovic et al., 2018). From the results in Table 6-B, we can see that our self-supervised mixed-modal jigsaw pre-training matches or improves on ImageNet performance for each of the FG-SBIR variants tested.

The effect of Sinkhorn Iterations l In practice, there is a trade-off on selecting the value of



(a) 3x3 jigsaw puzzle

(b) 4x4 jigsaw puzzle

(c) 5x5 jigsaw puzzle

Figure 6.5: Jigsaw solver success rate vs. Sinkhorn iterations once trained under l. Patch success rate and Instance success rate refer to the percentage of the shuffled patches that are correctly ordered and the percentage of the instances where all patches within are perfectly recovered respectively. Note that since it's practically infeasible to test all possible permutations of one sample, for each subfigure, we generate one mix-modality shuffling strategy for each input and apply it to all x-axis values.

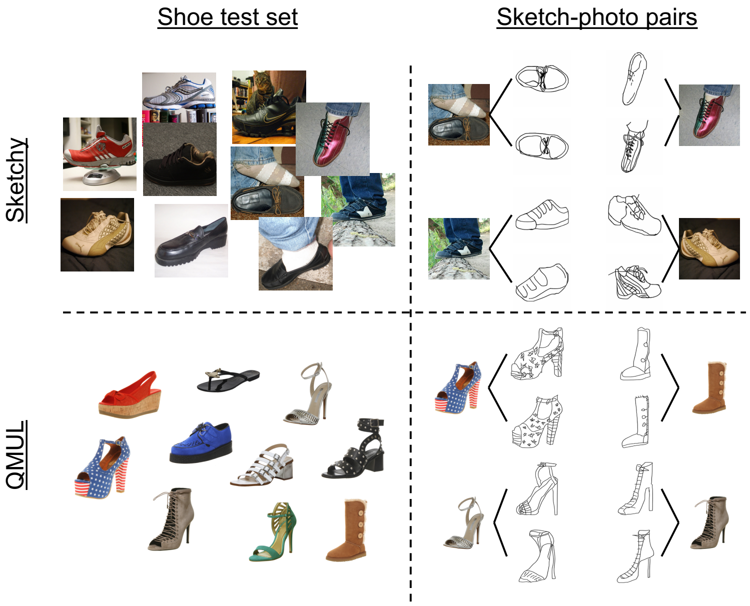


Figure 6.6: Illustrations of our product-level FG-SBIR dataset and the existing general-purpose counterpart, Sketchy (Sangkloy et al., 2016).

l: if it is too small, then the resultant assignment matrix will be far from a true permutation one, while when it's unhelpfully big, the optimisation becomes harder as the gradients vanished accordingly. In Figure 6.5, we show how jigsaw solver reacts to the linear slicing of different values ranging from 1 to l. The following observations can be made: (i) Generally, the jigsaw model saturates when the number is approaching l, with few exceptions that best performance is gained halfway (Figure 6.5(c)). (ii) For many settings after one round of Sinkhorn normalisation, the jigsaw performance already reaches to a reasonable level. This implies that even if we apply l times of Sinkhorn iteration during training, the model only improve the solving success marginally, but continue to pre-train a better model. (iii) Despite failing to get instances perfectly un-shuffled, e.g., less than 1% on 5×5 puzzle, the solver can consistently get a large number of patches right. (iv) Different jigsaw granularities corresponds to very different scales of jigsaw success rates, in a stark contrast with that on FG-SBIR (Figure 6.4 (a)), where little difference is witnessed as long as the granularity exceeds 2x2.

Caveat: SBIR Dataset Flavours We note that thus far the superiority of our jigsaw pre-training is validated when applied to *product*-level FG-SBIR benchmarks because this is where FG-SBIR is most likely to be applied. Here we consider two other type of datasets: The

Dataset	**Methods**		
	Ours$^{shoe\_4\times4}$	Ours$^{shoe\_4\times4}$/ImageNet	ImageNet
Sketchy	53.45%± 0.28	51.86%± 0.17	**60.26%± 0.16**
Flickr15k	27.23%± 0.81	24.03%± 0.84	**44.15%± 0.30**

Table 6-C: Performance comparison on coarser-grained SBIR datasets. Values reported on Sketchy (Sangkloy et al., 2016) and Flickr15k (Bui et al., 2017) are measured with Acc@1 and mAP respectively.

Flickr15k (Bui et al., 2017) benchmark for *category-level* SBIR (i.e., the goal is to retrieve any instance of a particular category rather than one specific instance), and Sketchy (Sangkloy et al., 2016), with sketch-photo paired data covering 125 real-world object categories. We follow the standard splits for these benchmarks, and evaluate our Jigsaw pre-training approach vs. the standard ImageNet pre-training in Table 6-C. We can see that our Jigsaw strategy is not effective for these benchmarks, and direct ImageNet pre-training clearly leads to the best results. To understand why, we show in Figure 6.6 the test set photos of the shoe category in Sketchy and a random 10 shoe photos in QMUL_Shoe_V2. It can been seen : (i) *Pose* and *shape* play pivotal roles in matching for sketchy, rather than fine-grained details in product-level FG-SBIR. This lesser pose variability in QMUL_Shoe_V2 contributes to the poor transferability to Sketchy. (ii) Sketchy and Flickr15k images have complicated backgrounds, unlike the white-background product images. Pre-training on product photos thus is unsurprisingly ineffective in teaching a model to deal with complex backgrounds required for Sketchy and Flickr15k. In these cases ImageNet pre-training is understandably more appropriate.

6.7 Summary

This chapter introduced a new mixed-modal jigsaw self-supervised pre-training strategy for FG-SBIR with a novel solver. We showed that the proposed method outperforms the conventional ImageNet pre-training stage. This strategy generalises well across categories, and furthermore leads to FG-SBIR models with better cross-category generalisation properties. We hope this pre-training strategy can become the norm for future FG-SBIR work, as well as be adopted by other cross-modal retrieval/recognition tasks.

Chapter 7

Conclusion and Future Work

Because of the belief that seeing can be better explained by drawing and the practical availability of relevant large-scale datasets that makes a model to be evaluated of statistical significance, sketch-related researches have flourished in recent years. In this thesis, we focus on the problem of FG-SBIR and have described a number of data-driven deep learning approaches to effectively solve it. We have dedicated to making each approach to address a critical part unidentified in the prior work. This is achieved by delving deep into the unique visual characteristics of the human sketch domain and exploit it to define and devise novel frameworks for better instance-level sketch-photo matching. Overall, this thesis has pushed the frontier of FG-SBIR research not only from benchmarking perspective (as in all chapters) but also in a way to cater for the potential needs in the practical adoption: The contour-detail factorisation study in Chapter 4 may inspire a new human sketching interface by separately tracing and recording the contours and details part. Chapter 5 alleviates the additional complexity of a deployed FG-SBIR system – the model does not have to recollect data and be re-trained to give reasonable responses to the user's sketch queries from object categories it has not seen before. The behaviours of self-supervised Jigsaw pre-training in Chapter 6 shows that we can cut the category tie between model pre-training and fine-tuning stage. This is of particular interest to FG-SBIR system developers:

because a pre-training solution of universal category wisdom is possible, efforts can finally be focused on fine-tuning stage only.

However, while this thesis has concretely demonstrated its efficacy for FG-SBIR, the problem stills remains largely open and some possible improvements building on our current models are summarised here: (i) rather than posing the cross-domain generation task as an auto-encoder In Chapter 3, we can re-formulate it to variational auto-encoder (Kingma and Welling, 2013) or adversarial auto-encoder (Makhzani et al., 2015) to make the embedding space tractable and thus better regularised. (ii) The better performance of soft assignment strategy in Chapter 5 indicates our assumption that one sketch is represented by one visual trait descriptor (VTD) only is over-constrained. Relaxing the freedom to multiple VTDs and learning a dynamic weighting combination between them seems to be more reasonable. (iii) Chapter 6 shows that ideal self-supervised pre-training task depends on the downstream target task and Jigsaw prevails for FG-SBIR. A better pre-training model is expected by tuning the distribution over self-supervision objectives compared to picking any one of them or using them all with uniform weight. Simple meta learning technique like (Alex and Johnn, 2018) should be an effective way to start.

Below we discuss several potential future directions for FG-SBIR:

On-the-fly Fine-grained SBIR Two barriers still exist that hinders the practicality of FG-SBIR — the time taken to draw a complete sketch, and the drawing skill shortage of the user. Interactive FG-SBIR tackles this problem by aiming to conduct retrieval at every stroke drawn as opposed to the requirement of a complete sketch. This in practice can also enhance the user's sketching patience and simultaneously adapt their drawing behaviours because one can see the immediate result of each incremental rendering.

Editing-based Fine-grained SBIR Thus far, the human sketches for FG-SBIR all assume a drawing process starting from a white canvas. However, in practical scenarios, we may often have an initial image and want to edit upon. This requires different model capabilities where both the impact of the initial image should be considered (its difference to the target image is often very small) and much finer-grained differentiation is required (the few strokes of edited subtle part should be fully reflected in the retrieval).

Standardised FG-SBIR Dataset Collection Previous FG-SBIR datasets are collected mainly via two ways: crowd-sourcing platform like Amazon Mechanical Turk (e.g., Sketchy (Sangkloy et al., 2016) or private recruitment (e.g., QMUL-V1 datasets (Yu et al., 2016)). The former has the advantage of effortlessly scaling the size of the dataset while the latter prevails on quality control. Nevertheless, there still lacks a systematic study on quantitatively evaluating the faithfulness of collected sketch-photo pairs and how it will affect the FG-SBIR model learning. Because unlike other human-involved annotation processes, for example drawing bounding boxes in object detection (Lin et al., 2014) or locating key points (Cao et al., 2018) in pose estimation, the subjectivity in the process of sketching based on a mental image is particularly high. This can constitute a major casual factor for data-driven deep models .

Appendix A

Author's Publications

Conference Papers

1. **K. Pang**, Y. Song, T. Xiang and T. Hospedales, "Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval", *In Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, 2017.

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

3. K. Li, **K. Pang**, J. Song, Y. Song, T. Xiang, and T. Hospedales, "Universal Sketch Perceptual Grouping", *In Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

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

5. P. Xu, Y. Huang, T. Yuan, **K. Pang**, Y. Song, T. Xiang, and T. Hospedales, Z. Ma and J. Guo, "Sketchmate: Deep Hashing for Million-Scale Human Sketch Retrieval", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,

Salt Lake City, Utah, USA, 2018.

6. **K. Pang***, K. Li*, Y. Yang, T. Hospedales, T. Xiang, and Y. Song, "Generalising Fine-Grained Sketch-Based Image Retrieval", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, 2019.

7. **K. Pang**, Y. Yang, T. Hospedales, T. Xiang, and Y. Song, "Solving Mixed-modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, USA, 2020.

.

Journal Papers

1. K. Li, **K. Pang**, Y. Song, T. Hospedales, T. Xiang and H. Zhang, "Synergistic Instance-Level Subspace Alignment for Fine-Grained Sketch-Based Image Retrieval", *IEEE Transactions on Image Processing (TIP)*, 26(12), pp.5908-5921.

2. K. Li, **K. Pang**, J. Song, Y. Song, T. Xiang, T. Hospedales and H. Zhang, "Universal Sketch Perceptual Grouping", *IEEE Transactions on Image Processing (TIP)*, 28(7), pp.3219-3231.

Bibliography

Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

Nichol Alex and Schulman Johnn. Reptile: A scalable meta-learning algorithm. `https://blog.openai.com/reptile/`, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002.

Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in neural information processing systems (NIPS)*, 2016.

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Tu Bui and John Collomosse. Scalable sketch-based image retrieval using color gradient features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2015.

Tu Bui, L Ribeiro, Moacir Ponti, and John Collomosse. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. *Computer Vision and Image Understanding (CVIU)*, 2017.

Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 2018.

Xiaochun Cao, Hua Zhang, Si Liu, Xiaojie Guo, and Liang Lin. Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *Proceedings of the 18th ACM International Conference on Multimedia (ACMMM)*, 2010.

Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Abdolah Chalechale, Golshah Naghdy, and Alfred Mertins. Sketch-based image matching using angular partitioning. *IEEE Transactions on Systems, Man, and Cybernetics*, 2004.

Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yin Chans, Zhibin Lei, Daniel P Lopresti, and Sun-Yuan Kung. Feature-based approach for image retrieval by sketch. In *Multimedia Storage and Archiving Systems II*, 1997.

Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Transactions on Graphics (TOG)*, 2009.

Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Gabriela Csurka. *Domain Adaptation in Computer Vision Applications*. Springer, 2017.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

Alberto Del Bimbo and Pietro Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1997.

Jia Deng, Wei Dong, R Socher, and Li Jia Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 2010.

Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on Graphics (TOG)*, 2012.

Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marac Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2011.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2009.

Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, 1985.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems (NIPS)*, 2013.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 2016.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems (NIPS)*, 2017.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. In *Advances in neural information processing systems (NIPS)*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems (NIPS)*, 2017.

Kyoji Hirata and Toshikazu Kato. Query by visual example. In *International Conference on Extending Database Technology*, 1992.

Rui Hu and John Collomosse. A performance evaluation of gradient field HOG descriptor for sketch-based image retrieval. *Computer Vision and Image Understanding (CVIU)*, 2013.

Rui Hu, Tinghuai Wang, and John Collomosse. A bag-of-regions approach to sketch-based image retrieval. In *IEEE International Conference on Image Processing (ICIP)*, 2011.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Toshikazu Kato, Takio Kurita, Nobuyuki Otsu, and Kyoji Hirata. A sketch retrieval method for full color image database-query by visual example. In *Proceedings of the 11th IAPR International Conference on Pattern Recognition*, 1992.

Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy M. Hospedales, Tao Xiang, and Honggang Zhang. Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval. *IEEE Transactions on Image Processing (TIP)*, 2017b.

Ke Li, Kaiyue Pang, Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Honggang Zhang. Universal sketch perceptual grouping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

Yi Li, Yi-Zhe Song, Timothy M Hospedales, and Shaogang Gong. Free-hand sketch synthesis with deformable stroke models. *International Journal of Computer Vision (IJCV)*, 2017c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017b.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems (NIPS)*, 2016.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems (NIPS)*, 2017c.

Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Stanislaw Matusiak, Mohamed Daoudi, Thierry Blu, and Olivier Avaro. Sketch-based images database retrieval. In *International Workshop on Multimedia Information Systems*, 1998.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS)*, 2013.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

Sarthak Parui and Anurag Mittal. Similarity-invariant sketch-based image retrieval in large databases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yonggang Qi, Yi-Zhe Song, Tao Xiang, Honggang Zhang, Timothy Hospedales, Yi Li, and Jun Guo. Making better use of edges via perceptual grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

R Kumar Rajendran and Shih-Fu Chang. Image retrieval with sketches and compositions. In *2000 IEEE International Conference on Multimedia and Expo*, 2000.

Antti Rasmus, Harri Valpola, and Tapani Raiko. Lateral connections in denoising autoencoders support supervised learning. *arXiv preprint arXiv:1504.08215*, 2015.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*, 2015.

Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, 2018.

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Jose M Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *IEEE International Conference on Image Processing (ICIP)*, 2014.

Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems (NIPS)*, 2016.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. In *ACM Transactions on Graphics (TOG)*, 2016.

Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. Deeppermnet: Visual permutation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitionn (CVPR)*, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Omar Seddati, Stéphane Dupont, and Saïd Mahmoudi. Quadruplet networks for sketch-based image retrieval. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Mastering sketching: adversarial augmentation for structured prediction. *ACM Transactions on Graphics (TOG)*, 2018.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy Hospedales, and Xiang Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Finegrained image retrieval: the text/sketch input dilemma. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017a.

Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017b.

Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Giorgos Tolias and Ondrej Chum. Asymmetric feature maps with application to sketch based retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research (JMLR)*, 2016.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Yongxin Yang and Timothy M Hospedales. A unified perspective on multi-domain and multi-task learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Yongxin Yang and Timothy M. Hospedales. Multivariate regression on the grassmannian for predicting novel domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch-based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems (NIPS)*, 2014.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Sketch-a-net that beats humans. *arXiv preprint arXiv:1501.07873*, 2015.

Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106*, 2017a.

Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. SketchX! - Shoe/Chair fine-grained SBIR dataset. `http://sketchx.eecs.qmul.ac.uk`, 2017b.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann LeCun. Stacked what-where autoencoders. In *Proceedings of the International Conference on Learning Representations Workshop Track (ICLR Workshop Track)*, 2015.

Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)*, 2014.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Larry Zitnick and Piotr Dollar. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.