International Journal of Computer Vision (2021) 129:484–500
https://doi.org/10.1007/s11263-020-01382-3

Fine-Grained Instance-Level Sketch-Based Image Retrieval

Qian Yu[1,2] · Jifei Song[2] · Yi-Zhe Song[2] · Tao Xiang[2] · Timothy M. Hospedales[2,3]

Received: 29 March 2019 / Accepted: 5 September 2020 / Published online: 30 September 2020

Abstract

The problem of fine-grained sketch-based image retrieval (FG-SBIR) is defined and investigated in this paper. In FG-SBIR, free-hand human sketch images are used as queries to retrieve photo images containing the same object instances. It is thus a cross-domain (sketch to photo) instance-level retrieval task. It is an extremely challenging problem because (i) visual comparisons and matching need to be executed under large domain gap, i.e., from black and white line drawing sketches to colour photos; (ii) it requires to capture the fine-grained (dis)similarities of sketches and photo images while free-hand sketches drawn by different people present different levels of deformation and expressive interpretation; and (iii) annotated cross-domain fine-grained SBIR datasets are scarce, challenging many state-of-the-art machine learning techniques, particularly those based on deep learning. In this paper, for the first time, we address all these challenges, providing a step towards the capabilities that would underpin a commercial sketch-based object instance retrieval application. Specifically, a new large-scale FG-SBIR database is introduced which is carefully designed to reflect the real-world application scenarios. A deep cross-domain matching model is then formulated to solve the intrinsic drawing style variability, large domain gap issues, and capture instance-level discriminative features. It distinguishes itself by a carefully designed attention module. Extensive experiments on the new dataset demonstrate the effectiveness of the proposed model and validate the need for a rigorous definition of the FG-SBIR problem and collecting suitable datasets.

Keywords Fine-grained · Sketch understanding · Image retrieval · Cross-modality · Deep learning

1 Introduction

Existing image retrieval paradigms are still dominated by methods that use text or exemplar images as input (Krizhevsky and Hinton 2011; Moulin et al. 2014; Johnson et al. 2015; Noh et al. 2017). Since the main applications of image retrieval is to find specific object instances (e.g., a particular shoe worn by a pedestrian that one just saw on the street), the two modalities have different strengths and weaknesses: textual queries are easy to obtain (just involving typing some words), but often unable to accurately describe the visual appearance of the object instance (e.g., it can be a tall order for a non-fashion-expert to describe exactly what a shoe looks like); in contrast, an image is worth a thousand words, so if a picture of that object can be obtained, instance-level image retrieval is made much easier; however, taking a photo could be difficult or even not possible (e.g., it would be generally considered to be rude to take a photo of a stranger's shoe on the street).

Due to the proliferation of touch-screen devices, only very recently has sketch-based image retrieval (SBIR) started to return as a practical form of retrieval (Eitz et al. 2010, 2011; Lin et al. 2013; James et al. 2014; Wang et al. 2015; Bui et al. 2016, 2018; Zhang et al. 2018). Compared with text, sketches are incredibly intuitive to humans and have been used since

Communicated by Patrick Perez.

Q. Yu, J. Song have contributed equally to this work.

✉ Qian Yu
qianyu@buaa.edu.cn

Jifei Song
j.song@qmul.ac.uk

Yi-Zhe Song
y.song@surrey.ac.uk

Tao Xiang
t.xiang@surrey.ac.uk

Timothy M. Hospedales
t.hospedales@ed.ac.uk

[1] Beihang University, Beijing, China

[2] SketchX, CVSSP, University of Surrey, Surrey, UK

[3] University of Edinburgh, Edinburgh, UK

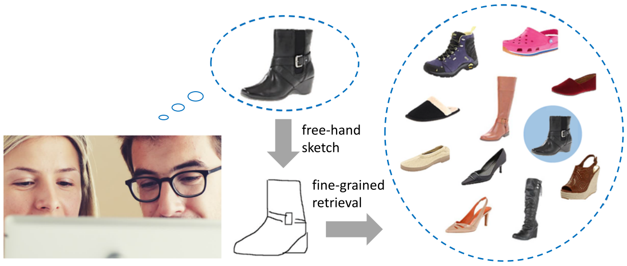

Fig. 1 Free-hand sketch is ideal for fine-grained instance-level image retrieval

pre-historic times to conceptualise and depict visual objects (Marr 1982; Landay and Myers 2001). Furthermore, a unique characteristic of sketches in the context of image retrieval is that they offer inherently fine-grained visual descriptions – a sketch speaks for a 'hundred' words. Importantly, compared with photos, it is also much easier to produce: it can be obtained almost anywhere and anytime based on a mental recollection of the object instance.

However, existing SBIR studies mainly focus on retrieving images of the same category as a depicted sketch (Eitz et al. 2010, 2011; Hu et al. 2010; Cao et al. 2011, 2010; Wang et al. 2010; Hu et al. 2011; Lin et al. 2013; James et al. 2014; Wang et al. 2015; Hu and Collomosse 2013; Bui et al. 2016, 2018), thus not exploiting the real fine-grained strength of SBIR for instance-level retrieval. This oversight pre-emptively limits the practical value of SBIR since the text is often a simpler form of input when only category-level retrieval is required. E.g., one would rather type in the word "shoe" to retrieve the target object rather than sketching a shoe. Existing commercial image search engines already do a great job of category-level image retrieval based on text queries. In contrast, it is when aiming to retrieve *a particular shoe* that sketching may be preferable than elucidating a long textual description of it. Figure 1 illustrates an application scenario of using free-hand sketch for fine-grained image search.

This paper investigates the problem of *fine-grained* SBIR (FG-SBIR) at *instance-level*, opening a new research direction for human free-hand[1] sketch analysis in computer vision. Specifically, we consider that since in most application scenarios, especially those online shopping scenarios, when a user resorts to sketch as the means of query input, s/he has already known what category the object instance belongs to. Importantly, the gallery photo images to be searched against have also been organised into specific object categories to limit the search space (e.g. one would search in the shoe section of a shopping website for shoes). FG-SBIR is thus mainly about matching object instance of a given category across the sketch and photo domains. In contrast, the existing attempts of defining FG-SBIR either confuse instance to pose retrieval (Li et al. 2014) or do not clearly separate the category-level and instance-level SBIR problems (Sangkloy et al. 2016).

As a cross-domain instance-level retrieval problem, FG-SBIR faces a number of challenges. First, sketches and photos are from inherently heterogeneous domains – sparse black and white line drawings versus dense colour pixels. Second, FG-SBIR requires representation of fine-grained (dis)simila-rities of sketches and photos. However, different people often have very different drawing styles and abilities meaning that sketches come with varied levels of deformation and abstraction. This makes the instance-level matching between sketches and photos a non-trivial problem. Specifically, given a query sketch, there are often many visually similar candidate photos in the gallery while the true match may only differ subtly in some localised object parts with other wrong matches (as shown in Fig. 2a); on the other hand, for a specific object, such as the shoe and chair in Fig. 2b, the sketches drawn by different people could have very different appearance due to the drastically different drawing styles, varying levels of abstraction and the loss of colour and texture information in the sketch domain.

Last but not least, FG-SBIR is challenged by having only scarce benchmarking datasets. A FG-SBIR dataset is harder to obtain than a category-level one because each sketch should have an instance-level corresponding photo. Importantly such dataset should capture sufficient variability of the human drawing styles so that any model learned from the data can have a chance to capture the large domain differences and drawing style variability.

In this paper, these challenges are effectively addressed. Specifically, to address the lack of data problem, we introduce a large-scale fine-grained SBIR database, named *QMUL FG-SBIR database*. This database has a number of unique characteristics that make it suitable for tackling the newly-defined FG-SBIR problem: (1) It is large-scale, consisting of 4 datasets, including 3116 photos and 8721 sketches in total belonging to two categories (shoe and chair). (2) It is designed carefully to reflect the challenges in real-world application scenarios: multiple sketches are collected for each photo to capture the drawing style variations, and the sketches are collected using different input devices and from users with different levels of familiarisation with SBIR. (3) Extensive data annotation are provided: in addition to the sketch-photo pairs, 32,220 human triplet annotations are obtained, which provide valuable training data for learning the subtle differences between two similar photos to a given query sketch.

Existing FG-SBIR models focus primarily on closing the semantic gap between the two domains whilst only partially addressing or completely ignoring the challenge of capturing instance-level discriminative features. Specifically,

[1] *Free-hand* sketch in this work refers to sketches drawn by amateurs based on their mental recollection. Specifically, we assume that before a human draws a sketch, (s)he has seen a reference object instance, but does not have the object or a photo at hand while drawing.

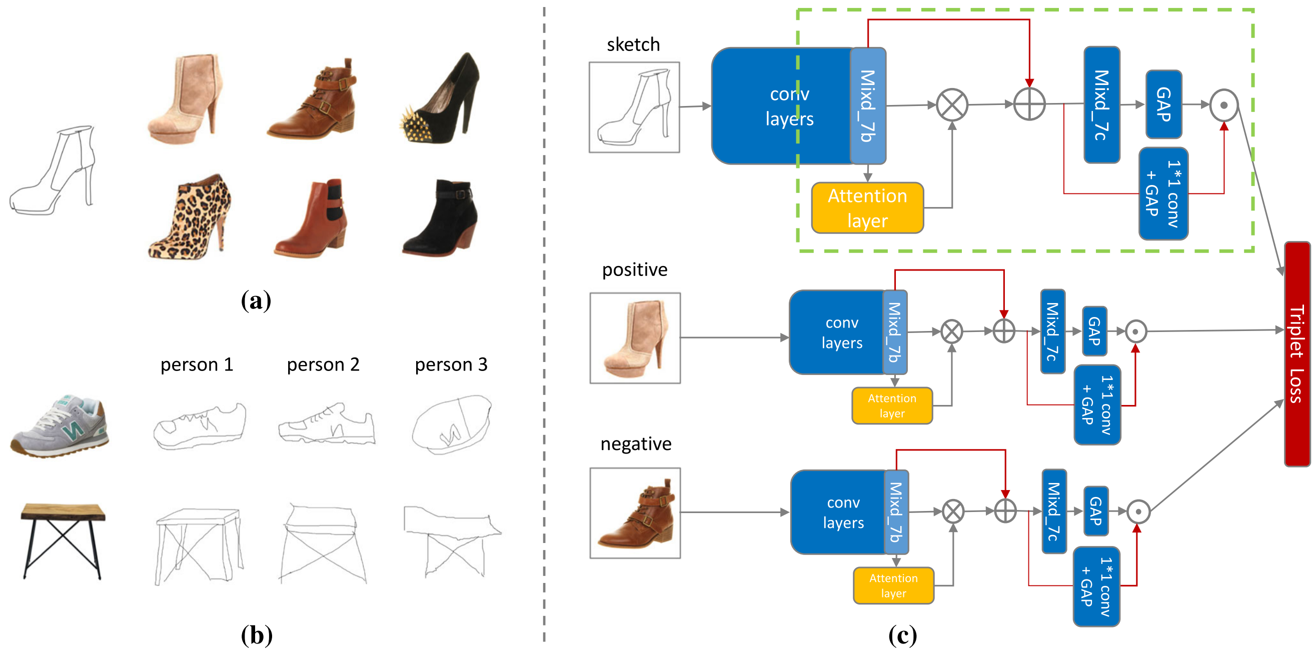

Fig. 2 **a** An example of a query sketch with several visually similar candidate photos; **b** human free-hand sketches have various abstraction and deformation levels and are visually very different from the photo images containing the same object instance—the bottom three sketches were drawn by three people depicting the shoe in the same photo in the top; **c** architecture of our proposed model

a FG-SBIR model, such as (Sangkloy et al. 2016) and the previous version (Yu et al. 2016), adopts a multi-branch deep convolutional neural networks (CNNs). Each domain has a corresponding branch which consists of multiple convolutional/pooling layers followed by fully connected (FC) layers. The final FC layer is used as input to pairwise verification or triplet ranking losses to align the domains. However, recent efforts (Gatys et al. 2015; Mahendran and Vedaldi 2015) on visualising what each layer of a CNN actually learns show that higher-layers of the network capture more abstract semantic concepts but not fine-grained details, motivating fine-grained recognition methods to work with convolutional feature maps instead (Lin et al. 2015). After the FC layers, the fine-grained detail is gone and cannot be recovered. Thus existing deep FG-SBIR models are unable to tell apart visually similar photos based on subtle differences.

In this paper, we introduce spatial-semantic attention modelling in deep FG-SBIR to perform effective instance-level matching of sketch and photo given the large domain gap and variable sketch drawing styles. The architecture of the proposed model is shown in Fig. 2c. Although it is still essentially a multi-branch CNN, there are a number of crucial differences to existing models. First, we introduce attention modelling in each branch of the CNN so that computation for representation learning is focused on specific discriminative local regions rather than being spread evenly over the whole image. Due to the large misalignment between the sketch and photo domains, directly taking the attended feature map as input to the subsequent layers of the network is too sensitive to misalignment. We thus introduce a shortcut connection architecture (Szegedy et al. 2015; He et al. 2016) to link the input directly to the output of the attention module so that a noisy attention mask would not derail the deep feature computation completely, resulting in robust attention modelling. Second, we keep both coarse and fine semantic details through another shortcut block to connect the attended feature map with the final FC layer before feeding it to the loss.

Our contributions are as follows: (1) For the first time, the problem of fine-grained instance-level image retrieval using free-hand sketches is defined and addressed. (2) We contribute a large-scale fine-grained sketch database, *QMUL FG-SBIR*, with extensive ground truth annotations, in the hope that it will inspire research efforts on solving this challenging problem. (3) We propose a cross-domain attention model for FG-SBIR, providing insights for this task. Extensive experiments on the new dataset show that the proposed model significantly outperforms the state-of-the-art alternatives. We also demonstrate that the contributed dataset is more suitable than the existing ones [e.g., Sketchy(Sangkloy et al. 2016)] for studying the FG-SBIR problem.

2 Related Work

Category-level SBIR Most existing SBIR studies (Eitz et al. 2010, 2011; Hu et al. 2010; Cao et al. 2011, 2010; Wang et al. 2010; Hu et al. 2011; Lin et al. 2013; James et al. 2014; Hu and Collomosse 2013; Bui et al. 2016; Liu et al. 2017a; Bui et al. 2018; Zhang et al. 2018) focus on category-level SBIR. In this task, a sketch is used to retrieve photos of the same category. For example, given *any* sketch of a cat, e.g., a cat face or a standing cat, *any* photo containing a cat is deemed a match for it. (Bui et al. 2016, 2018) proposed staged-training strategy and investigated different base networks for category-level SBIR. (Liu et al. 2017a) recently introduced deep hashing into category-level SBIR, making it practical for commercial applications by speeding up the retrieval process on large-scale datasets. (Zhang et al. 2018) further presented a generative hashing model which can learn a mapping for sketches that the distribution is indistinguishable from that of photos using an adversarial loss. Although they demonstrated the effectiveness of the proposed model in both category-level and instance-level settings, the focus of this work is still on efficiency.

Another similar task is sketch-based 3D shape retrieval, like (Wang et al. 2015), which uses the sketch to retrieve 3D shapes. And (Collomosse et al. 2017) explored style-constrained sketch search over a diverse domain of images with different visual aesthetics. However, these related tasks are also limited to category-level.

Fine-grained Instance-Level SBIR There is a smaller but growing number of studies addressing the fine-grained instance-level SBIR problem. However, the definitions of what FG-SBIR entails differ from ours. The first recognisably FG-SBIR problem was proposed in (Li et al. 2014). Nevertheless, in this work a gallery photo is considered to match a query sketch if the objects depicted are in a similar arrangement, i.e. in the same *pose* and depicted with similar viewpoint and zoom parameters. However, there is no requirement for the photo to depict the same *instance*, as is our focus. Fine-grained SBIR was studied in the context of multiple categories in a concurrent work (Sangkloy et al. 2016) with ours, which also introduced the first large-scale FG-SBIR dataset called Sketchy. This work is thus closely related, but it also has a number of vital differences. Firstly, like (Li et al. 2014), pose and viewpoint are the dominant cues for matching in the Sketchy dataset. Thus it does not test a model's capability to differentiate fine-grained *instance-level* details, which are more subtle and challenging to match across domains, and would underpin a practical commercial application. Secondly, while retrieval among a gallery of multiple categories is an interesting challenge, it is not very relevant to a practical application: where a user would more reasonably use conventional keyword tools to specify the category and then perform FG-SBIR to find a specific instance within that category. We perform detailed analysis contrasting the nature of the Sketchy FG-SBIR task and our instance-level FG-SBIR task in Sect. 5. Finally, the FG-SBIR models proposed in (Sangkloy et al. 2016) have a Heterogeneous network architecture, whilst in this study we show that far superior performance can be obtained with a Siamese architecture. Besides, a recent work (Radenovic et al. 2018) whose focus is shape matching, shows that with careful designing, a model that is trained on edge maps can generalize well on sketches. However, it involves a series of data pre-processing and hard mining, and the performance still has a gap with our proposed method.

Other SBIR works like Sketch2Photo (Chen et al. 2009) and AverageExplorer (Zhu et al. 2014), use sketch in addition to text or colour cues for image retrieval. (Zhu et al. 2014) further investigates an interactive process, in which each user 'edit' indicates the traits to focus on for refining retrieval. For now, we focus on non-interactive black & white sketch-based retrieval and leave these extensions to future work.

Another relevant task is instance-level image retrieval which aims at retrieving all images that contain the same object instance as the query image. Works like (Gordo et al. 2017; Radenović et al. 2018) achieve impressive performance on landmark datasets such as Oxford 5k (Philbin et al. 2007) and Paris 6k (Philbin et al. 2008). Different from the task studied in this work, instance-level image retrieval is conducted among natural images. Therefore, the domain gap between query and target images is not as significant as in SBIR.

Fine-grained SBIR Datasets One of the key barriers to fine-grained SBIR research is the lack of large-scale benchmark datasets. There are free-hand sketch datasets designed for sketch recognition, the most commonly used being the TU-Berlin 20,000 sketch dataset (Eitz et al. 2012); there are also many photo datasets such as PASCAL VOC (Everingham et al. 2010) and ImageNet (Deng et al. 2009). Therefore, with few exceptions (Eitz et al. 2011; Hu and Collomosse 2013), most existing SBIR benchmarks were created by combining overlapping categories of sketches and photos from existing databases, which means that only category-level SBIR is possible. The fine-grained dataset contributed in (Li et al. 2014) was created by selecting similar-looking sketch-photo pairs from the TU-Berlin and Pascal VOC datasets, thus without the guarantee that those pairs contain the same object instances. For each of 14 categories, there are 6 sketches and 60 images—much smaller than ours, and too small to apply state-of-the-art deep learning techniques. For specific domains such as face, large-scale datasets exist such as the CUHK Face Sketches (Wang and Tang 2009). However, those forensic sketches were drawn by trained artists rather than the general public with variable drawing abilities.

The only existing SBIR dataset to our knowledge is the Sketchy dataset proposed in the concurrent work (Sangkloy et al. 2016), which is larger than ours including 74,425 sketches and 12,500 gallery photos spanning 125 categories collected from social media sites such as Flickr. In each category, there are 100 object instances, and each instance has 1 photo and 5 or more corresponding sketches. Compared to our QMUL FG-SBIR database, there are several key differences: (i) Sketchy involves multiple object categories. As mentioned early, this does not reflect the real-world application scenarios where the object category is given. Obviously it could be re-purposed for within-category SBIR. However, in that case the number of object instances per-class (100) is too small (the largest dataset in our database contains 2000 instances). (ii) It includes many categories not appropriate for FG-SBIR, like pizza, banana and pear etc. Instances in such categories can only be distinguished by colour or texture information, which cannot or hardly be reflected in sketches. (iii) Pose is a dominant factor for distinguishing instances of many categories. For example, for most animal classes, such as cat, bat and cow, the most distinct feature of an instance is its pose, particularly given the small instances numbers. Thus FG-SBIR essentially degenerates to a pose detection task for these categories. In contrast, in our database all photos are from online shopping websites where pose and background are most often identical, forcing the FG-SBIR models to focus on detecting subtle fine-grained intrinsic visual properties (e.g., having a higher heel or an additional buckle for a shoe) for matching. More detailed analysis with visual illustrations on the differences between Sketchy and the proposed FB-SBIR database can be found in Sect. 5.6.

Attention modelling Visual attention models have been studied extensively in a wide range of vision problems including image caption generation (Xu et al. 2015; Lu et al. 2016), VQA (Fukui et al. 2016; Nam et al. 2016), image classification (Mnih et al. 2014; Sermanet et al. 2014; Xiao et al. 2015) and particularly fine-grained image recognition (Sermanet et al. 2014; Xiao et al. 2015). Various types of attention models exist. Soft attention is the most commonly used one because it is differentiable thus can be learned end-to-end with the rest of the network. Most soft-attention models learn an attention mask which assigns different weights to different regions of an image. Alternatively, the spatial transformer network (Jaderberg et al. 2015) generates an affine transformation matrix which locates the discriminative region. Different from soft attention, hard attention models only indicate one region at each time. A hard attention model is not differentiable so it is typically learned using reinforcement learning. Beyond the soft- and hard-attention models which are developed based on convolutional neural networks, (Vaswani et al. 2017) introduces a novel architecture, *Transformer*, which solely based on attention mechanism but does not involve convolutions. It is brought up for language translation. Interestingly, there is no prior SBIR (both category-level and instance-level) work that models attention, perhaps because conventional attention models deployed in a cross-domain match problem assume pixel-level alignment; they thus become ineffective when this assumption is invalid as in the case of SBIR. Our attention model is specifically designed for FG-SBIR in that it is robust against spatial misalignment through the shortcut connection architecture.

Shortcuts and layer fusion in deep learning The shortcut architecture used in both the attention module and the coarse-fine fusion block in our model serves to fuse multiple layers at different depths. Fusing different CNN layers in the model output has been exploited in many problems such as edge detection [e.g., (Ren 2008; Xie and Tu 2015)], pose estimation [e.g., (Newell et al. 2016)] and scene classification [e.g., (Gong et al. 2014; Yang and Ramanan 2015; Liu et al. 2017b)]. The motivation is typically multi-scale (coarse to fine) fusion rather than attended-unattended feature map fusion, as in our first shortcut block.

Various shortcut connection architectures have been successfully deployed in a number of widely used CNNs including GoogLeNet (Szegedy et al. 2015) and ResNet (He et al. 2016). Our shortcut connection architecture is similar to that of the residual block in ResNet (He et al. 2016). However, instead of making the network deeper, we use it in the attention module to make the attention module output robust against noisy attention mask caused by cross-domain feature misalignment, as well as in the final CNN output layer to preserve both coarse and fine-grained information in the learned representation.

A preliminary version of this work was published in (Yu et al. 2016; Song et al. 2017). Compared with the earlier studies, there are several key differences: (i) This work contributes a much bigger fine-grained SBIR dataset, including two collection settings, reflecting different application scenarios. (ii) The proposed triplet ranking model with a different base network and modified pre-training strategy significantly outperforms the model in (Yu et al. 2016; Song et al. 2017) on our new dataset. (iii) A detailed comparison between our new dataset and the existing one, i.e., Sketchy is conducted to provide more insights in this task.

3 The QMUL FG-SBIR Database

3.1 Overview

Our QMUL FG-SBIR database consists of 3116 photos and 8721 sketches belonging to two categories (shoes and chairs). Each category has two datasets giving a total of four datasets. For a given category, the two datasets are collected under different settings to reflect different people's drawing abilities and styles as well the drawing devices typically used in a real-world application.

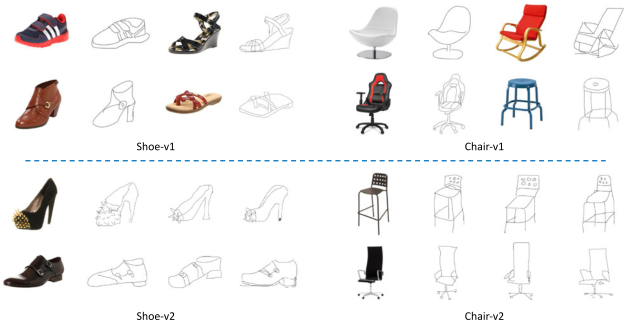

Fig. 3 Example sketch-photo pairs in QMUL FG-SBIR database. Top: V1, Bottom: V2. Each photo in V1 has one corresponding sketch while photos in V2 have 3 or more corresponding sketches drawn by different persons. It is clear to see that (i) human sketches are abstract and iconic in nature, (ii) drawing abilities and styles are different among participants

More specifically, the first setting (V1) is in-house collection in a controlled environment, where all volunteers were students familiar with SBIR, and sketched on our provided tablets. The second setting (V2) is uncontrolled under which sketches were collected on Amazon Mechanical Turk (AMT). On this platform, diverse workers used various input devices resulting in a much greater diversity of sketches, and we purposefully recorded multiple (3+) sketches from different workers per object instance to reflect such drawing diversity.

We denote the shoes/chairs collected under the first setting as Shoe/Chair-v1 and the latter as Shoe-v2 and Chair-v2. Figure 3 shows the example sketches and photos of V1 and V2. Note that the V1 datasets are exactly those introduced in the preliminary version of this work (Yu et al. 2016) and used in (Song et al. 2017). Generally, sketches in V1 are visually better (more detailed and truthfully reflective of the subtle details in the corresponding photos) than those in V2. This is due to the controlled input device and greater familiarity of volunteers with SBIR. V1 thus simulates a future scenario where people have got used to drawing on touch-screens, while V2 is more representative of the contemporary general public. A detailed comparison of the four datasets is listed in Table 1.

3.2 Data Collection

Collecting Photo Images Because our database is designed for instance-level retrieval, the photo images should cover the instance-level variability of the visual appearance of the corresponding object category. To this end, for Shoe-v1, we selected 419 representative photo images from UT-Zap50K (Yu and Grauman 2014) representative shoes of different types including boots, high-heels, ballerinas, formal and informal shoes. While for Shoe-v2, we exhaustively collected the photos from the largest shoes on-line shop in the UK, *OFFICE (http://www.office.co.uk/)*. We filtered out the shoes with the same style but different colours because sketches in this work do not contain colour information. The final selection consists of 2,000 shoe photos. For Chair-v1, we searched three on-line shopping websites, including IKEA, Amazon and Taobao, and selected 297 chair product photos of varying types and styles. We collected another 400 photos from several furniture websites, including *Argos* (http://www.argos.co.uk/) and *MADE* (http://www.made.com/) for Chair-v2. Note that the photo images of V1 and V2 are collected from different websites, thus they are mostly disjoint. To further remove duplicated photos with the same object, we extracted image features from photos of V1 and preliminary V2 (pre-V2) using AlexNet, and searched for the top-10 nearest neighbors from pre-V2 for each photo of V1. Then we manually checked and removed the duplicated photos.

Table 1 Comparison of V1 (Shoe/Chair-v1) and V2 (Shoe/Chair-v2) datasets

Items	V1	V2
#Photos	419/297	2000/ 400
#Sketches	419/297	6730/1275
#Volunteers	82	599
Devices	Tablets	Tablets, smartphones, mouse

Collecting Sketches The second step is to use the collected photo images to generate corresponding sketches.

Shoe/Chair-v1: The sketches of V1 are collected in a controlled environment. 22 out of 82 recruited volunteers are assigned to sketch the images and the rest 60 for data annotation (to be detailed later). We showed one shoe/chair image to a volunteer on a tablet for 15 seconds, then displayed a blank canvas and let the volunteer sketch the object he/she just saw using his/her fingers on the tablet. None of the volunteers has any art training but they are familiar with the sketch collection process. Each photo has one corresponding sketch, so there are 419 and 297 sketches for shoes and chairs respectively.

Shoe/Chair-v2: We collected sketches from AMT for V2. We showed one shoe/chair photo to a worker for 5 seconds, then the photo disappeared and the worker needed to sketch the object on a blank canvas. The worker can re-check the photo for multiple times but the canvas is cleared after reviewing the photo. Because people have different drawing abilities and styles, the same object would be depicted differently by different sketchers. To explore this, we collected 3 or more sketches for each photo from different workers. Finally, we collected 6730 and 1275 sketches for shoes and chairs, respectively from 599 workers. In addition, we recorded the worker-ID and the stroke-level temporal information (Yu et al. 2015) of collected sketches for future study.

3.3 Data Annotation

The data will be used to train a FG-SBIR model that is able to find the most similar photos to a query sketch. The photo-sketch pair correspondence already provides some annotation that could be used to train a pairwise verification model (Chopra et al. 2005). However, for fine-grained analysis it may be possible to learn a stronger model by using a ranking-based loss, provided that we have detailed annotation of the similarity ranking of candidate photos with respect to a given query sketch. To explore whether this denser but more subjective and noisy annotation improves performance, we also collect annotations in the form of human similarity judgments. Given limited resources, we only collected such annotations for V1 datasets. For V2 datasets, we always used the true match as positive while the rest as negative when training a ranking-loss based model.

Ranking annotation for FG-SBIR is not straightforward. Asking a human annotator to rank all 419 shoe photos given a query shoe sketch would be an error-prone task. This is because humans are bad at list ranking, especially given the factor that many shoes look very alike; but they are better at individual forced choice judgments or pairwise ranking which has been employed in (Gygli et al. 2013; Jiang et al. 2013). Therefore, instead of requiring a global ranking annotation, a much more manageable triplet ranking task is designed for the annotators. Specifically, each triplet consists of one query sketch and two candidate photos. The task is to determine which of the two candidate photos is more similar to the query sketch. Exhaustively annotating all possible triplets is also out of the question due to the extremely large number of possible triplets. We therefore selected only a subset of the triplets and obtained the annotations through the following three steps:

1. **Attribute Annotation:** We first defined an ontology of attributes for shoes and chairs based on the existing UT-Zap50K attributes (Yu and Grauman 2014) and product tags on on-line shopping websites. We selected 21 and 15 binary attributes for shoes and chairs respectively. 60 volunteers annotated all 1,432 images (i.e., both sketches and photos) with ground-truth attribute vectors. All these attribute annotations will be provided in our database.
2. **Generating Candidate Photos for Each Sketch:** Next we selected 10 most-similar candidate images for each sketch in order to focus our limited amount of gold-standard fine-grained annotation effort. In particular, we combined the attribute vector with a deep feature vector (the fc7 layer features extracted using Sketch-a-Net (Yu et al. 2015)) and computed the Euclidean distance between each sketch and image. For each query sketch, we took the top 10 closest photo images to the query sketch as candidates for annotation.
3. **Triplet Annotation:** To provide triplet annotations for the $(419 + 297) \cdot 10 \cdot 9/2 = 32,220$ triplets generated in the previous step, each volunteer was presented with one sketch and two photos at a time. They were then asked to indicate which photo is more similar to the sketch. Each sketch has $10 \cdot 9/2 = 45$ triplets and three people annotated each triplet. We merged the three annotations by majority voting to clean up some human errors.

4 Methodology

4.1 Overview

The architecture of the proposed model is illustrated in Fig. 2c. It is a Siamese network with three CNNs, taking a query sketch, a positive photo and a negative photo as the input respectively. The positive-negative relation can be defined by the matching relationship, e.g., if the true match photo is the positive, any false match can be used as the negative. Alternatively, if the sketches and photos are annotated explicitly by similarity, relative similarity ordering can be used as supervision information. The CNNs extract deep features from the three input images and feed them to a triplet ranking loss to enforce the ranking order (positive should be closer to the query than the negative using the extracted feature). Specifically, for a given triplet $t = (s, p^+, p^-)$, its loss is defined as:

$$L_t = \max(0, \Delta + D(f_\theta(s), f_\theta(p^+)) - D(f_\theta(s), f_\theta(p^-))) \tag{1}$$

where Δ is a margin between the positive-query distance and negative-query distance. If the two photos are ranked correctly with a margin of distance Δ, then this triplet will not be penalised. Otherwise the loss is a convex approximation of the $0-1$ ranking loss which measures the degree of violation of the desired ranking order specified by the triplet.

After the training, the model can be used to do the instance-level SBIR in the inference stage. For a given query sketch s and a set of M candidate photos $\{p_j\}_{j=1}^M \in \mathscr{P}$, we construct the ranking score as:

$$R(s, p_j) = -D\left(f_\theta(s), f_\theta(p_j)\right) \tag{2}$$

where $R(s, p_j)$ denotes the ranking score between the query sketch and the candidate photo. $f_\theta(s)$ and $f_\theta(p_j)$ are the learned feature embedding. $D(\cdot, \cdot)$ is the same distance function used in the training stage, i.e., Euclidean distance. Note that the features should be normalized before computing the distance. Given a query sketch, the candidate photo in the gallery with the highest ranking score will be retrieved as the most similar photo.

The proposed spatial-semantic attention model has two key components: (1) a residual attention module and (2) coarse-fine feature fusion. Compared with the preliminary version (Song et al. 2017), the CNN base net is changed to InceptionV3 as its performance in sketch recognition reaches 80.23%, which surpasses previous state-of-the-art. In addition, we simplified the data preprocessing step by using RGB photos as input instead of edge maps extracted from photos. This is because InceptionV3 can effectively model both sketches and photos, and this also facilitates the usage of existing datasets for pre-training, which will be detailed in Sect. 4.4.

4.2 Attention Modelling

A soft attention paradigm is adopted. Given a feature map computed at any convolutional layer of a CNN, a soft attention module will take it as input and generate an attention mask. This mask is then used to re-weight the input feature map to get an attended feature map which is fed into the next layer of the network. In our model, the attention module is added to the output of layer Mixed_7b of the CNN in each branch (orange boxes in Fig. 2c).

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

$$s_{i,j} = F_{att}(f_{i,j}; \mathbf{W}_a), \quad \alpha_{i,j} = softmax(s_{i,j}), \tag{3}$$

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

$$f_{i,j}^{att} = \alpha_{i,j} \odot f_{i,j}. \tag{4}$$

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

$$\mathbf{f}_s^{att} = \mathbf{f} + \alpha \odot \mathbf{f}, \tag{5}$$

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

4.3 Coarse-Fine Fusion

Although the final attended feature map $\mathbf{f}_s^{att}$ is spatially aware and attentive to fine-grained details, these tend to be lost going through multiple subsequent fully connected layers, defeating the purpose of introducing attention modelling. To keep both the coarse and fine-grained information, a shortcut connection architecture is again employed here. Specifically, we fuse the attended feature map $\mathbf{f}_s^{att}$ with the output of the final layer (Mixed_7c) $\mathbf{f}^{Mixed\_7c}$ to form the final feature representation $\mathbf{f}^{final}$ before it is fed into the loss layer. A simple concatenation operation is used to fuse the two features. Before the fusion, we first reduce the dimension of the attended feature from 2,048D to 512D via a 1 × 1 convolutional layer and then do global average pooling (GAP).

4.4 Training Strategy

Preprocessing for Photo Branch There is still no consensus on whether we should use extracted edge map or the original RGB photo as the input representation for the photo branch. Some works (Li et al. 2014; Yu et al. 2016) suggest that processing photos to edges can reduce the domain gap between sketch and photo domains, thus alleviating the burden of aligning the two domains. In contrast, some other works (Sangkloy et al. 2016) argue that some information is lost in the edge map extraction process which could be useful but can never be recovered; they thus advocate no preprocessing of the photo input. We also select the original photo as the input rather than the detected edge map since the original photo contains more detailed information.

Pretraining Strategy Given the limited amount of training data, and the fine-grained nature of the instance-level SBIR task, training a good deep ranker is extremely challenging. In practice, it requires carefully designing of the training strategy (Yu et al. 2016; Sangkloy et al. 2016). As our base network is pretrained for the ImageNet-1K object category classification task, it is already suitable for category-level recognition. Turning attention to the goal of FG-SBIR, we initialise our three branch triplet network with three ImageNet-1K pre-trained InceptionV3 and then further pretrain it on Sketchy database which has a Hybrid category-instance SBIR task, thus getting closer to our final single

Table 2 The training/testing splits of QMUL FG-SBIR and Sketchy databases

Item	Sketchy	QMUL FG-SBIR			
		Shoe-v1	Chair-v1	Shoe-v2	Chair-v2
# Photos	12,500	419	297	2000	400
# Sketches	74,425	419	297	6730	1275
# Categories	125	1	1	1	1
# Triplet annotations	–	32,220		–	
# Skeches per photo	5+	1		3+	
# Sketches/Photos (train)	65,064/11,250	304/304	200/200	6051/1800	951/300
# per-category	∼520/90	304/304	200/200	6051/1800	951/300
# Sketches/Photos (test)	6312/1250	115/115	97/97	679/200	324/100
# per-category	∼50/10	115/115	97/97	679/200	324/100

category SBIR task. While training on Sketchy, both category classification and triplet loss are applied since there are 125 categories in the dataset. The classification loss ensures category separability. We generate triplets *within* each class: For each triplet, the positive is the true match photo for the anchor sketch while the negative is randomly selected from the other photos of the same class. The Sketchy pretrained model can be used for fine-grained instance-level retrieval directly: We will show that it does generalise to our fine-grained datasets. However, it is advantageous to further fine-tune the triplet model specifically for the target category. In our case, this means that the Sketchy pre-trained model is finally fine-tuned on the training split of our contributed QMUL FG-SBIR database.

5 Experiments and Results

5.1 Datasets

Three datasets are used in our experiments: (i) **QMUL FG-SBIR database** is our newly collected database, consisting of 4 datasets spanning 2 classes (shoes and chairs). There are 419 and 297 sketch-photo pairs in Shoe-v1 and Chair-v1; and 2000 photos and 6730 sketches in Shoe-v2, and 400 photos and 1275 sketches in Chair-v2. The detailed training/testing split is listed in Table 2. (ii) The QMUL Handbag dataset was introduced in the earlier version of this work (Song et al. 2017). It contains 568 sketch/photo pairs, and was collected to make the retrieval task more challenging, since handbags may exhibit more complex visual patterns and shapes than shoes and chairs. Note that we did not expand this handbag dataset due to limited resources. (iii) **Sketchy data-base** (Sangkloy et al. 2016) is the largest existing SBIR database, including 74,425 sketches and 12,500 gallery photos spanning 125 categories. In each category, there are 100 photos and 5 or more corresponding sketches for each photo. In Sect. 5.6, a detailed comparison between Sketchy and our dataset is provided to show their differences. In our experiments, this dataset is used for pre-training.

5.2 Implementation Details

Our method is implemented on the Tensorflow platform. Adam optimiser is applied to optimise the loss function. The initial learning rate is set to 0.0001 and the batch size is 16. The margin Δ (see Eq. 1) is set to be 0.3. During training, we do random cropping and flipping for data augmentation. ImageNet followed by Sketchy provide pre-training for QMUL FG-SBIR. Human triplet annotations are used as supervision when training on the Shoe-v1 and Chair-v1. For the others for which no triplet annotations are available, given a sketch, the true-match photo is used as positive while the rest as negative. We implemented two models, one is the basic model without employing attention module while the other does (i.e., our full model). Similar to (Song et al. 2017), the attention module in the full model consists of 2 convolutional layers, both with kernel size 1 × 1. The dimension of $\mathbf{f}^{final}$ is 2,560D.

Multi-view Testing: Similar to the multi-view testing procedure used in (Krizhevsky et al. 2012) for the image classification task, we also use multi-view testing for fine-grained SBIR. In details, we first crop patches sized 224×224 located on left-upper, left-down, right-upper, right-down and center of a given query sketch image, as well as their horizontal reflections. We then apply the trained model to extract features for each version of the query, calculate the ranking score between the corresponding patches in the gallery photo images, and then calculate the average score for evaluation of the retrieval performance.
Evaluation Metrics: For performance evaluation, the *retrieval accuracy* of the true-match photo for a given query sketch is used. We quantify this by computing

the cumulative matching accuracy at various ranks—so acc.@K is the percentage of sketches whose true-match photos are ranked in the top K retrieval results. This is the most commonly used evaluation metric for image retrieval tasks, corresponding to an application scenario where the goal is simply to find a specific item/image as quickly as possible without requiring the user to browse pages after pages of retrieved candidate images.

5.3 Competitors

We compare our proposed method (Sect. 4) with several popular shallow and deep baselines.

Shallow Baselines The venerable HOG feature is a standard feature that has long been applied to sketch-recognition (Li et al. 2015) and SBIR (Li et al. 2014; Hu and Collomosse 2013). We first extracted HOG feature from image patches, then directly concatenate the image's HOG features to form the dense HOG feature (576D). After that, we feed the dense HOG feature to train a rankSVM as in (Prosser et al. 2010) (denoted as *Dense-HOG+rankSVM*).

Deep Baselines Two groups of deep baselines are considered. In the first group, we extract deep feature using the deep Sketch-a-Net model (Yu et al. 2017) and InceptionV3 (Szegedy et al. 2016a). We then follow the same RankSVM pipeline to train a retrieval model (*ISN Deep+RankSVM* and *InceptionV3+rankSVM*). Furthermore, a recent method (Li et al. 2017) is also compared which further aligns the GoogleNet extracted features across the two domains with fine-grained subspace learning (*Triplet GoogleNet +Subspace*). In the second group, FG-SBIR models are learned end-to-end. *Triplet SN* and *Triplet Att. SN* are proposed in the earlier version of this work (Yu et al. 2016; Song et al. 2017). Similar to our model, *Triplet SN* also adopts a three-branch Siamese network architecture with a triplet ranking loss, and *Triplet Att.SN* adds an attention module and an HOLEF loss to improve the performance further. The newly proposed model differs from these two mainly in two aspects: (1) The base network in each branch is Sketch-a-Net (Yu et al. 2015) which is purposely built for sketch analysis whilst the general-purpose InceptionV3 (Szegedy et al. 2016b) is employed in our model. (2) The input for the photo branch is edge map extracted from the original photo whilst no such pre-processing step is taken in our model. On the Sketchy database, we also include the model presented in (Sangkloy et al. 2016) (*Triplet GoogleNet*) for comparison. This model uses a heterogeneous architecture with triplet ranking loss and is shown to be the best performing model among a number of variants on Sketchy in (Sangkloy et al. 2016). Two of those variants in (Sangkloy et al. 2016) that use pairwise losses are compared in our experiments.

Human Baselines In order to give an intuition about how challenging the FG-SBIR task is, we also collect data on human performance for each of our new databases. For each query sketch, the participants browsed all the gallery photos and selected the most similar one as the human retrieval result. We provided a two-stage selection procedure to help humans to perform this task more accurately. The participants can select multiple putative matching photos at the first round, and then focus on making a decision among these similar photos in the second round. We also include the human baseline on the Sketchy database, as reported in (Sangkloy et al. 2016).

5.4 Comparisons Against State-of-the-Art

Table 3 shows the performance of our proposed models on QMUL FG-SBIR dataset and the handbag dataset against the baselines. We make the following observations: (i) Our model achieves the highest accuracy at acc.@1 and acc.@10, often significantly outperforming the second best model. Especially on Chair-V2, the proposed model surpasses human performance by a noticeable margin. (ii) Dense-HOG outperforms the deep feature extracted from Sketch-a-Net, i.e., ISN Deep, on all datasets. This can be explained by that the sketches and photos on our contributed datasets are pose aligned. But when the general-purpose InceptionV3 is used, the handcrafted features fare much worse. (iii) End-to-end learned models are in general stronger compared with models with separate feature extraction and retrieval modelling steps. (iv)We also compare with the results of (Radenovic et al. 2018) on Shoe/Chair-v1. The focus of (Radenovic et al. 2018) is shape matching, but it also conducts experiments on fine-grained SBIR and achieves impressive performance. However, ours still outperforms theirs by a noticeable margin.

Next, we conduct ablation studies on QMUL Shoe-v2 and Chair-v2 datasets.

5.5 Ablation Study

Contributions of Each Component We have introduced two novel components in our model: a residual attention module being robust to spatial misalignment and the coarse-fine fusion (CFF[2]) to combine the attended convolutional feature map with the final layer output. In order to evaluate the contributions of each component, we compare our full model (**Full**) with three stripped-down versions: baseline model with residual attention module only (**Base+attention**), base-

[2] Here 'CFF' refers to the operation of combining the feature map extracted from an earlier layer with the final layer output. This is different with the meaning in the preliminary version (Song et al. 2017) where it indicates both feature fusion and residual attention module.

Table 3 Comparative results against baselines

Dataset	Method	Acc.@1	Acc.@10
QMUL Shoe-v1	Dense-HOG + rankSVM	23.48%	73.91%
	ISN Deep + rankSVM	20.00%	62.61%
	InceptionV3 + rankSVM	48.70%	91.30%
	Triplet SN (Yu et al. 2016)*	52.17 %	92.17 %
	Triplet Att. SN (Song et al. 2017)	61.74 %	94.78 %
	DSM (with whitening) (Radenovic et al. 2018)	54.8%	92.2%
	Our model	**66.09**%	**94.78**%
	Human	76.52%	–
QMUL Chair-v1	Dense-HOG + rankSVM	59.79%	96.90%
	ISN Deep + rankSVM	47.42%	82.47%
	InceptionV3 + rankSVM	83.50%	**100.00**%
	Triplet SN (Yu et al. 2016)*	72.16 %	98.96 %
	Triplet Att. SN (Song et al. 2017)	81.44 %	95.88 %
	DSM (with whitening) (Radenovic et al. 2018)	85.6%	97.9%
	Our model	**91.75**%	**100.0**%
	Human	94.85%	–
QMUL Handbag	Dense-HOG + rankSVM	15.5%	40.48%
	ISN Deep + rankSVM	9.5%	44.1%
	InceptionV3 + rankSVM	28.6%	75.0%
	Triplet SN (Yu et al. 2016)*	39.9 %	82.1 %
	Triplet Att. SN (Song et al. 2017)	49.4 %	82.7 %
	DSM (with whitening) (Radenovic et al. 2018)	51.2%	85.7%
	Our model	**61.90**%	**89.29**%
	Human	50%	–
QMUL Shoe-v2	Dense-HOG + rankSVM	11.63%	48.01%
	ISN Deep + rankSVM	7.21%	34.02%
	InceptionV3 + rankSVM	30.78%	78.35%
	Triplet SN (Yu et al. 2016)*	30.93%	72.02%
	Triplet Att. SN (Song et al. 2017)	–	–
	Our model	**42.27**%	**82.18**%
	Human	49.50%	–
QMUL Chair-v2	Dense-HOG + rankSVM	29.32%	75.31%
	ISN Deep + rankSVM	11.73%	57.40%
	InceptionV3 + rankSVM	48.15%	86.73%
	Triplet SN (Yu et al. 2016)*	45.06%	86.42%
	Triplet Att. SN (Song et al. 2017)	–	–
	Our model	**69.14**%	**97.22**%
	Human	63.00%	–

Best results are highlighted in bold

'*' The results of Triplet SN (Yu et al. 2016) are the updated ones which are higher than the published ones due to parameter retuning. The other baseline results are copied from (Song et al. 2017) and (Radenovic et al. 2018)

line model with coarse-fine fusion (**Base+CFF**), and baseline without either (**Base**), i.e., a triplet ranking model whose base net is InceptionV3. Table 4 shows that the fusion component can bring noticeable improvement while combining two proposed elements, i.e., our final model, can achieve the best performance. An interesting thing is the proposed attention module can hurt the performance when working solely. This can be explained by random sampling strategy used in our experiments may result in candidate photos look very different from query sketches, as a result, it is hard for the

Table 4 Contributions of the different components

QMUL Shoe-V2	acc.@1	acc.@10
Base	36.08%	81.15%
Base + attention	31.37%	65.10%
Base + CFF	38.59%	82.92%
Full (our model)	**42.27%**	**82.18%**
QMUL Chair-V2	acc.@1	acc.@10
Base	57.41%	91.36%
Base + attention	57.10%	88.58%
Base + CFF	65.12%	94.75%
Full (our model)	**69.14%**	**97.22%**

Best results are highlighted in bold

Table 5 Comparison of using edge map and original photo as input to the photo branch of our model

QMUL Shoe-V2	acc.@1	acc.@10
edge map	35.20%	77.47%
RGB photo (ours)	**42.27%**	**82.18%**
QMUL Chair-V2	acc.@1	acc.@10
edge map	60.49%	95.99%
RGB photo (ours)	**69.14%**	**97.22%**

Best results are highlighted in bold

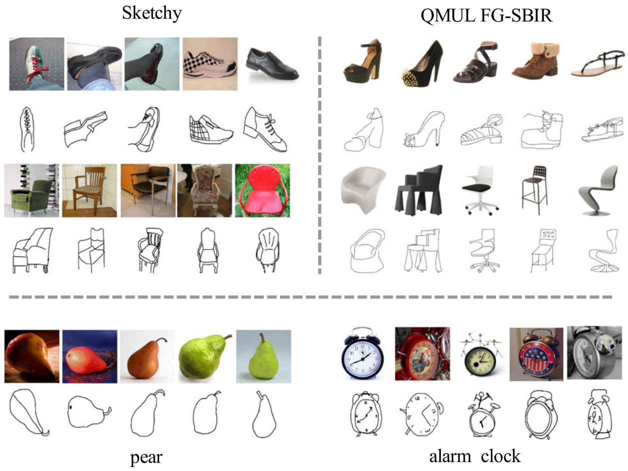

Fig. 4 Example photos and sketches of our QMUL FG-SBIR database and the Sketchy database. The upper part compares the shoes/chairs photos/sketches from these two databases. The bottom part shows some examples from the Sketchy database

attention module to learn a reliable attention mask. However, when working with our proposed fusion module, the attention module can be more effective because of the deep supervision introduced by the shortcut connection.

Effect of Training Strategy One of the key training strategy decision to make for FG-SBIR is whether to use the raw RGB photo image as input or its edge map in the hope that it can narrow the domain gap between sketch and photo. The former is adopted in our model. In this experiment, we investigate whether the edge-map extraction preprocessing step is necessary. Table 5 compares the performance of our model when the photo input is an edge map extracted using the edge detection method in (Zitnick et al. 2014) or RGB photo respectively. From the results, we can see that RGB photo always outperforms the corresponding edge map. This can not only simplify the data preprocessing but also make the model robust to photos with clustered background where edge maps cannot handle well.

5.6 Comparison of QMUL and Sketchy Databases

Both our QMUL FG-SBIR database and the Sketchy database are proposed for fine-grained image retrieval task. However, there are some vital differences (see Fig. 4). Firstly, our database was designed based on our model rigorous definition of FG-SBIR: it is a task for within-category instance-level photo retrieval using sketch as query. Consequently, each dataset in QMUL FG-SBIR only contains instances of the same category. Furthermore, the photos and sketches are well-aligned in terms of pose and viewpoint; thus the differentiating features of each shoe/chair is indeed identity (instance) related cues. In contrast, Sketchy is a multi-category dataset, so the model must perform both categorisation and instance recognition. Importantly, the instances in each category are also different in pose or viewpoints. As a result, pose or viewpoint becomes the key factor to distinguish photos and in many cases it shifts the task to a pose recognition task. As mentioned earlier, both category recognition and pose/viewpoint recognition are unnecessary in a real-world FG-SBIR application scenario.

Secondly, the instances within different categories in Sketchy exhibit varied levels of fine-grained recognisability. Some categories have distinct category-level characteristics (that makes them easy to recognise) but the intra-class differences are indistinguishable in sketch. For example, the pear category is included in Sketchy despite it is clearly unsuitable for instance recognition: distinguishing different pears is extremely difficult with photos and is impossible with sketches. Other categories may be relatively ambiguous at category-level, but their intra-class variation is large, making FG-SBIR relatively easy (e.g., across/within different animal categories).

Quantitative Results on Sketchy Dataset To demonstrate the difference between these two datasets further, we train a FG-SBIR model on Sketchy to show some quantitative

Table 6 Comparative results against baselines on Sketchy

Dataset	Method	Acc.@1	Acc.@10
Sketchy	Triplet GoogleNet + Subspace(Li et al. 2017)	45.27%	**98.20**%
	Triplet GoogleNet (Sangkloy et al. 2016)	37.10%	–
	Pairwise GoogleNet (Sangkloy et al. 2016)	27.36%	–
	Pairwise AlexNet (Sangkloy et al. 2016)	21.36%	–
	Triplet SN (Yu et al. 2016)	21.63%	67.60%
	Base	**59.55**%	96.56%
	Full (our model)	57.98%	**97.54**%
	Human (Sangkloy et al. 2016)	54.27%	–

Best results are highlighted in bold

results. FG-SBIR on Sketchy involves two tasks, including the category-level classification and instance-level image matching. We first retrain our proposed model (Sect. 4), i.e., the cross-domain attention model, plus a classification branch. Specifically, a classification layer is added on top of the final layer (Mixed_7c) of each branch, which accepts a feature vector (output of the last layer after global average pooling) as input. ImageNet provides data for model pre-training for Sketchy and the margin Δ is set as 0.1. The results are reported in Table 6. As we can see, the result of our proposed full model is slightly lower than **Base** model but significantly outperforms other baselines. This indicates that the retrieval task on Sketchy dataset is not affected much by the local attended parts, which verifies our assumption that the difference among different object instances lies in pose or orientation rather than identity related cues. Besides, in our experiments, we found the classification accuracy will decrease if we feed the same feature into the classification layer as for retrieval task, suggesting the local attended feature is not suitable for category-level classification. However, although the feature representations used for classification and retrieval task are different, i.e., fused feature for retrieval and final-layer output for classification, the proposed attention module harms classification accuracy.

Figures 5 and 6 show some retrieval results on our fine-grained datasets, the handbag dataset, and the Sketchy database respectively. It is clear to see that the proposed model can capture not only the holistic feature, like shape and posture, but also the pose-independent fine-details such as the buckle on a shoe or the pattern on the back of a chair.

6 Conclusion and Future Works

We have provided a rigorous definition of the FG-SBIR problem, elucidated its value, and contributed the largest single-category instance-level FG-SBIR benchmark. Through detailed evaluation of architectural choices, we have proposed a model that surpasses the state-of-the-art on all FG-SBIR benchmarks. A number of directions are worth further study. First, as shown in the Fig. 5, sketch has weaknesses as an input modality: It does not naturally reflect texture and colour information. Secondly, although sketches straightforwardly capture pose, position and shape, there is still large deformation, especially for deformable objects with complicated shape such as animals.

Given these weaknesses, the future FG-SBIR work should integrate the colour or texture information to improve the performance. In addition, from a practical point of view, other modalities such as text or attributes can be combined with sketch together. Besides, human feedback and human attention are another two important aspects for retrieval task. Given the fact that drawing a sketch is a dynamic process, it will be interesting to see if human attention can be detected from the drawing process and used to guide automated attention learning.

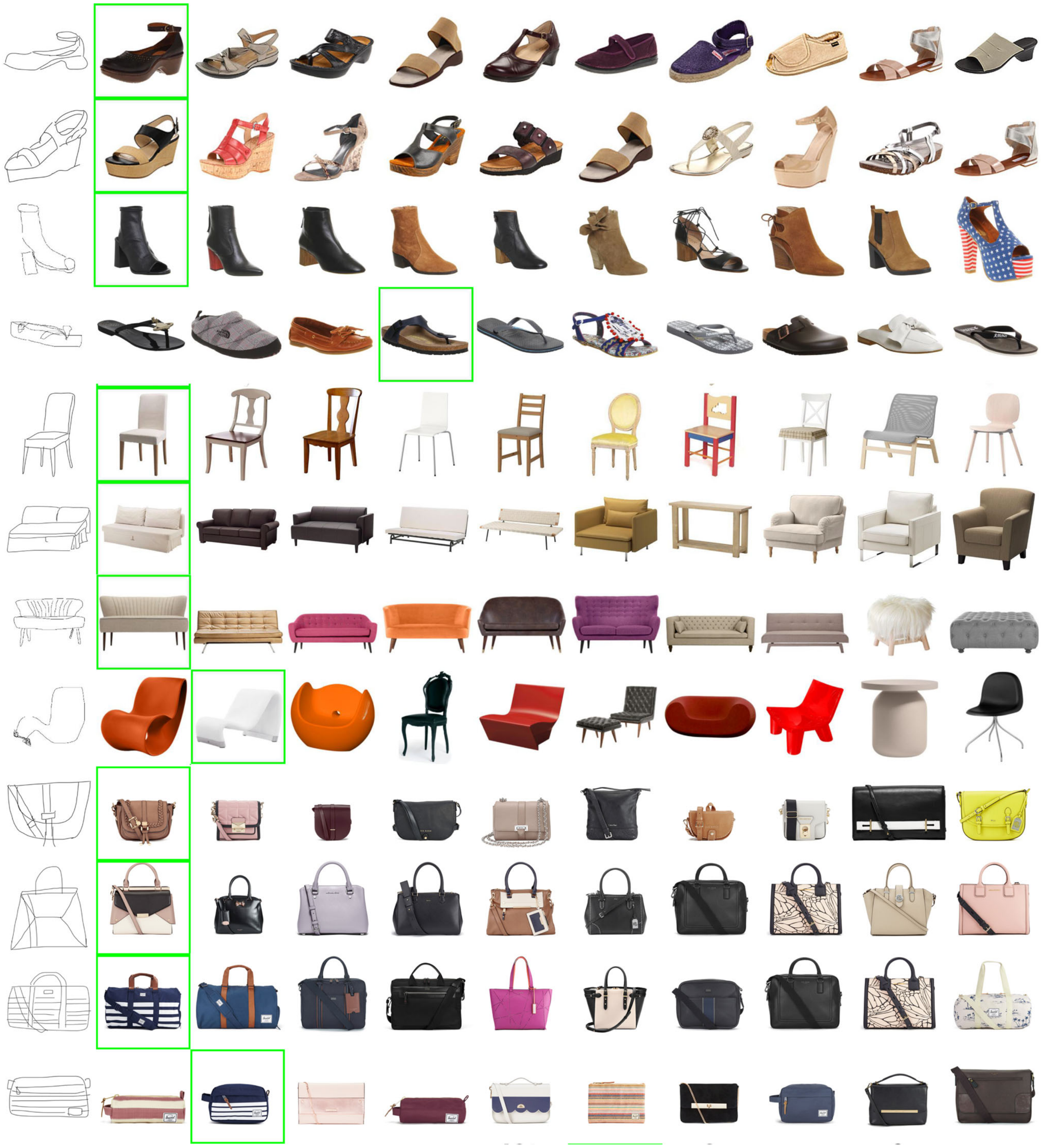

Fig. 5 Visualisation of the retrieved results on the new QMUL FG-SBIR database and the QMUL handbag dataset

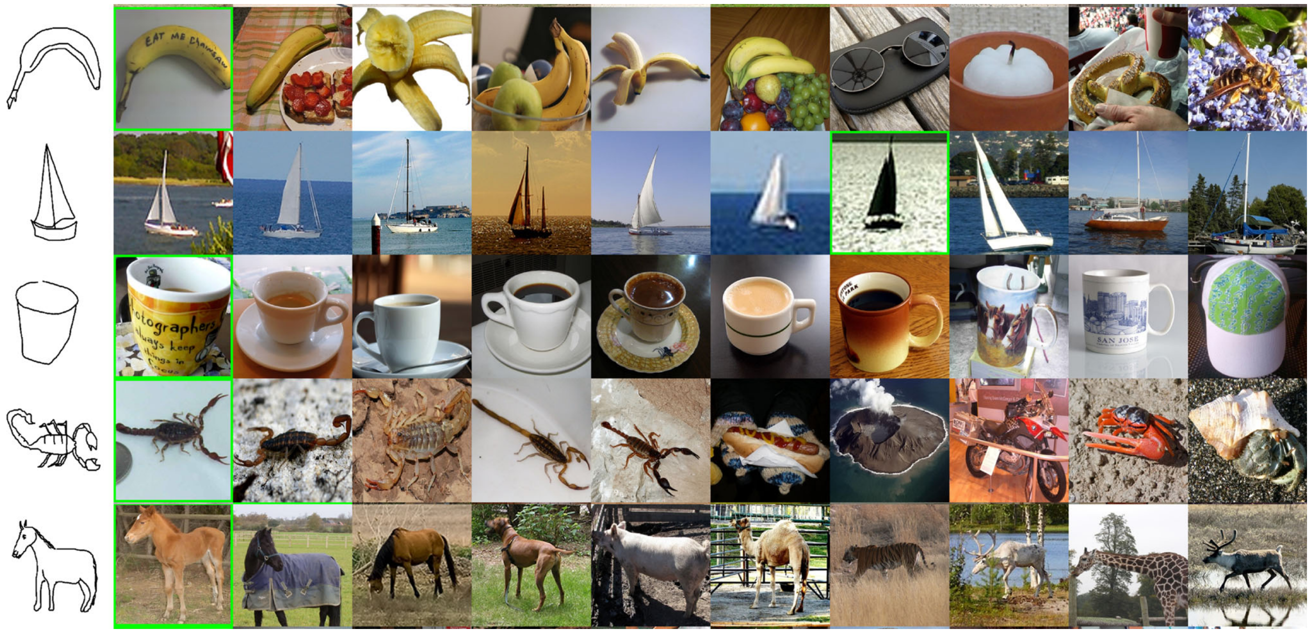

Fig. 6 Visualisations of retrieved results on Sketchy database

References

Bui, T., Ribeiro, L., Ponti, M., & Collomosse, J. (2016). Generalisation and sharing in triplet convnets for sketch based visual search. arXiv preprint arXiv:1611.05301.

Bui, T., Ribeiro, L., Ponti, M., & Collomosse, J. (2018). Sketching out the details: sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, *71*, 77–87.

Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., & Zhang, L. (2010). Mindfinder: interactive sketch-based image search on millions of images. In *International conference on multimedia*.

Cao, Y., Wang, C., Zhang, L., & Zhang, L. (2011) Edgel index for large-scale sketch-based image search. In *CVPR*.

Chen, T., Cheng, M. M., Tan, P., Shamir, A., & Hu, S. M. (2009). Sketch2photo: internet image montage. *ACM Transactions on Graphics (TOG)*, *28*, 1–10.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *IEEE computer society conference on computer vision and pattern recognitio*.

Collomosse, J., Bui, T., Wilber, M. J., Fang, C., & Jin, H. (2017). Sketching with style: visual search with sketches and aesthetic context. In *Proceedings of the IEEE international conference on computer vision*.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009) ImageNet: a large-scale hierarchical image database. In *CVPR*.

Eitz, M., Hildebrand, K., Boubekeur, T., & Alexa, M. (2010). An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, *34*(5), 482–498.

Eitz, M., Hildebrand, K., Boubekeur, T., & Alexa, M. (2011). Sketch-based image retrieval: benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization and Computer Graphics*, *17*(11), 1624–1636.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on Graphics (TOG)*, *31*, 1–10.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. CoRR, arXiv:1505.07376.

Gong, Y., Wang, L., Guo, R., & Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*.

Gordo, A., Almazan, J., Revaud, J., & Larlus, D. (2017). End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, *124*(2), 237–254.

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In *IEEE international conference on computer vision*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hu, R., & Collomosse, J. (2013). A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, *117*(7), 790–806.

Hu, R., Barnard, M., & Collomosse, J. (2010). Gradient field descriptor for sketch based retrieval and localization. In *IEEE international conference on image processing*.

Hu, R., Wang, T., & Collomosse, J. (2011). A bag-of-regions approach to sketch based image retrieval. In *IEEE international conference on image processing*.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in neural information processing systems*.

James, S., Fonseca, M., & Collomosse, J. (2014). Reenact: Sketch based choreographic design from archival dance footage. In *Proceedings of international conference on multimedia retrieval*.

Jiang, Y. G., Wang, Y., Feng, R., Xue, X., Zheng, Y., & Yang, H. (2013). Understanding and predicting interestingness of videos. In *AAAI*.

Johnson, J., Krishna, R., Stark, M., Li, L. J., Shamma, D., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *CVPR*.

Krizhevsky, A., & Hinton, G. E. (2011). Using very deep autoencoders for content-based image retrieval. In *European symposium on artificial neural networks*.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Landay, J. A., & Myers, B. A. (2001). Sketching interfaces: toward more human interface design. *IEEE Computer*, *34*(3), 56–64.

Li, Y., Hospedales, T., Song, Y. Z., & Gong, S. (2014). Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*.

Li, Y., Hospedales, T. M., Song, Y. Z., & Gong, S. (2015). Free-hand sketch recognition by multi-kernel feature learning. *Computer Vision and Image Understanding*, *137*, 1–11.

Li, K., Pang, K., Song, Y. Z., Hospedales, T. M., Xiang, T., & Zhang, H. (2017). Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval. *IEEE Transactions on Image Processing*, *26*(12), 5908–5921.

Lin, Y., Huang, C., Wan, C., & Hsu, W. (2013) 3D sub-query expansion for improving sketch-based multi-view image retrieval. In *Proceedings of the IEEE international conference on computer vision*.

Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 1449–1457).

Liu, L., Shen, F., Shen, Y., Liu, X., & Shao, L. (2017a). Deep sketch hashing: fast free-hand sketch-based image retrieval. arXiv preprint arXiv:1703.05605.

Liu, Y., Guo, Y., Lew, M. S. (2017b). On the exploration of convolutional fusion networks for visual recognition. In *International conference on multimedia modeling*.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2016). Knowing when to look: adaptive attention via a visual sentinel for image captioning. arXiv preprint arXiv:1612.01887.

Mahendran, A., & Vedaldi, A. (2015) Understanding deep image representations by inverting them. In *IEEE conference on computer vision and pattern recognition*.

Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*.

Moulin, C., Largeron, C., Ducottet, C., Géry, M., & Barat, C. (2014). Fisher linear discriminant analysis for text-image combination in multimedia information retrieval. *Pattern Recognition*, *47*(1), 260–269.

Nam, H., Ha, J. W., & Kim, J. (2016). Dual attention networks for multimodal reasoning and matching. arXiv preprint arXiv:1611.00471.

Newell, A., Yang, K., Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision*.

Noh, H., Araujo, A., Sim, J., Weyand, T., & Han, B. (2017). Large-scale image retrieval with attentive deep local features. In *IEEE international conference on computer vision*.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE conference on computer vision and pattern recognition*.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2008). Lost in quantization: improving particular object retrieval in large scale image databases. In *IEEE conference on computer vision and pattern recognition*.

Prosser, B. J., Zheng, W. S., Gong, S., Xiang, T., & Mary, Q. (2010). Person re-identification by support vector ranking. In *British machine vision conference*.

Radenovic, F., Tolias, G., & Chum, O. (2018). Deep shape matching. In *Proceedings of the European conference on computer vision*.

Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning cnn image retrieval with no human annotation. *TPAMI*, *41*(7), 1655–1668.

Ren, X. (2008). Multi-scale improves boundary detection in natural images. In *Proceedings of the European conference on computer vision*.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, *35*, 1–12.

Sermanet, P., Frome, A., & Real, E. (2014). Attention for fine-grained categorization. arXiv preprint arXiv:1412.7054.

Song, J., Yu, Q., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2017). Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. *IEEE conference on computer vision and pattern recognition*. arXiv:1409.4842.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016a). Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016b). Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *NeuIPS*.

Wang, X., & Tang, X. (2009). Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 1955–1967.

Wang, C., Li, Z., & Zhang, L. (2010). Mindfinder: image search by interactive sketching and tagging. In *Proceedings of the 19th international conference on world wide web*.

Wang, F., Kang, L., & Li, Y. (2015). Sketch-based 3D shape retrieval using convolutional neural networks. In *IEEE conference on computer vision and pattern recognition*.

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE conference on computer vision and pattern recognition*.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *IEEE international conference on computer vision*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: neural image caption generation with visual attention. In *International conference on machine learning*.

Yang, S., & Ramanan, D. (2015). Multi-scale recognition with DAG-CNNS. In *IEEE international conference on computer vision*.

Yu, A., & Grauman, K. (2014). Fine-grained visual comparisons with local learning. In *IEEE conference on computer vision and pattern recognition*.

Yu, Q., Yang, Y., Song, Y., Xiang, T., & Hospedales, T. (2015). Sketch-a-net that beats humans. In *BMVC*.

Yu, Q., Liu, F., Song, Y. Z., Xiang, T., Hospedales, T. M., & Loy, C. C. (2016). Sketch me that shoe. In *IEEE conference on computer vision and pattern recognition*.

Yu, Q., Yang, Y., Liu, F., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: a deep neural network that beats humans. *International Journal of Computer Vision*, *122*(3), 411–425.

Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., Tao Shen, H., & Van Gool, L. (2018). Generative domain-migration hashing for sketch-to-image retrieval. In *Proceedings of the European conference on computer vision (ECCV)*.

Zhu, J. Y., Lee, Y. J., & Efros, A. A. (2014). Averageexplorer: interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)*, *33*, 1–11.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: locating object proposals from edges. In *Proceedings of the European conference on computer vision*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.