Fine-Grained Sketch-Based Image Retrieval: The Role of Part-Aware Attributes

Ke Li[1&2] Kaiyue Pang[1&2] Yi-Zhe Song[2] Timothy Hospedales[2] Honggang Zhang[1]
Yichuan Hu[1]
[1]Beijing University of Posts and Telecommunications.
[2]School of Electronic Engineering and Computer Science Queen Mary University of London.

Abstract

We study the problem of fine-grained sketch-based image retrieval. By performing instance-level (rather than category-level) retrieval, it embodies a timely and practical application, particularly with the ubiquitous availability of touchscreens. Three factors contribute to the challenging nature of the problem: (i) free-hand sketches are inherently abstract and iconic, making visual comparisons with photos more difficult, (ii) sketches and photos are in two different visual domains, i.e. black and white lines vs. color pixels, and (iii) fine-grained distinctions are especially challenging when executed across domain and abstraction-level. To address this, we propose to detect visual attributes at part-level, in order to build a new representation that not only captures fine-grained characteristics but also traverses across visual domains. More specifically, (i) we propose a dataset with 304 photos and 912 sketches, where each sketch and photo is annotated with its semantic parts and associated part-level attributes, and with the help of this dataset, we investigate (ii) how strongly-supervised deformable part-based models can be learned that subsequently enable automatic detection of part-level attributes, and (iii) a novel matching framework that synergistically integrates low-level features, mid-level geometric structure and high-level semantic attributes to boost retrieval performance. Extensive experiments conducted on our new dataset demonstrate value of the proposed method.

1. Introduction

Sketches are intuitive and descriptive. They are one of the few means for non-experts to create visual content. As a query modality, they offer a more natural way to provide detailed visual cues than pure text. Closely coupled with the proliferation of touch-screen devices and availability of large scale free-hand sketch datasets [8], sketch-based image retrieval (SBIR) has gained tremendous application potential.

Traditional computer vision methods for SBIR mainly focus on category-level retrieval, where intra-category variations are neglected. This is not desirable, since if given a specific shoe sketch (e.g., high-heel, toe-open) as input, it would be pointless to retrieve an image that is indeed a shoe, but with different part semantics (e.g., a flat running shoe). Thus *fine-grained* SBIR is emerging as a way to go beyond conventional category-level SBIR, and fully exploit the detail that can be conveyed in sketches. By providing a mode of interaction that is more expressive than the ubiquitous browsing of textual categories, fine-grained SBIR is more likely to underpin practical commercial adoption of SBIR technology. Figure 1 contrasts our fine-grained SBIR with traditional category-level SBIR systems.

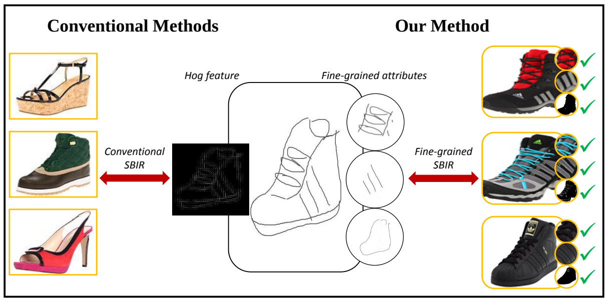


Figure 1. Conventional SBIR operates at *category*-level, but *fine-grained* SBIR requires more scrutiny at subtle details on an instance-level basis. We propose a part-aware learning approach to train our semi-semantic representations based on a new large-scale fine-grained SBIR dataset of shoes. (Best viewed in color.)

Fine-grained SBIR is challenging due to: (i) free-hand sketches are highly abstract and iconic, e.g., sketched objects do not accurately depict their real-world image counterparts. (ii) sketches and photos are from inherently heterogeneous domains, e.g., sparse black line drawings with white background versus dense color pixels, potentially with background clutter. (iii) *fine-grained* correspondence between sketches and images is difficult to establish especially given the abstract and cross-domain nature of the problem. Over and above all, there is lack of a purpose built *fine-grained* SBIR dataset to drive research, which is why we contribute a new FG-SBIR dataset to the community.

There exist significant prior work [32, 4, 16, 15, 17, 23] on retrieving images or 3d models based on sketches, typically with Bag Of Words (BOW) descriptors or advancements thereof. Although BOW approaches are effective and scalable, they are weak at distinguishing *fine-grained* variations as they do not represent any semantic information. Very recently, approaches to fine-grained SBIR have included DPM-based part modeling in order to retrieve objects in specific poses [21]. However, for practical SBIR in commercial applications, we are more interested in distinguishing subtly different object sub-categories rather than different poses. In a related line of work, fine-grained *attributes* have recently been used to help drive fine-grained image retrieval by identifying subtle yet semantic properties of images [7, 35]. Moreover, such attributes may provide a route to bridge the sketch/photo modality gap, as they are domain invariant if reliably detected (e.g., a high-heel shoe is 'high-heel' regardless if depicted in a photo or sketch). However, they suffer from being hard to predict due to spurious correlations [18]. In this paper we bring together attribute and part-centric modeling to decorrelate and better predict attributes, as well as provide two complementary views of the data to enhance matching.

We first define a taxonomy of 13 discriminative attributes commonly possessed by shoes, and acquire a large *fine-grained* SBIR dataset of free-hand shoe sketches with part-level attribute annotations. Based on this, we propose a part-aware SBIR framework that addresses the *fine-grained* SBIR challenge by identifying discriminative attributes and parts, and then building a synergistic representation based on them. Specifically, we first train strongly-supervised deformable part-based model (SS-DPM) to obtain semantic localized regions, followed by low-level features (i.e., HOG) extraction, geometric part structure extraction (mid-level) and semantic attribute prediction (high-level). We then use canonical correlation analysis (CCA) to get a robust subspace integrating all three views as our final feature representation. At retrieval time, we apply nearest neighbor matching to retrieve images most similar to the probe sketch. We demonstrate the superiority of our framework on *fine-grained* SBIR through in-depth comprehensive and comparative experiments.

The overall contributions of our work are:

- We propose a *fine-grained* SBIR shoe dataset with free-hand human sketches and photos, as well as fine-grained attribute annotations.
- We propose a part-aware paradigm that allows *fine-grained* attribute detection.
- We propose a synergistic low-level + mid-level + high-level feature representation that proves to be crucial to improve the performance of *fine-grained* SBIR.

2. Related Work

2.1. Sketch-based image retrieval

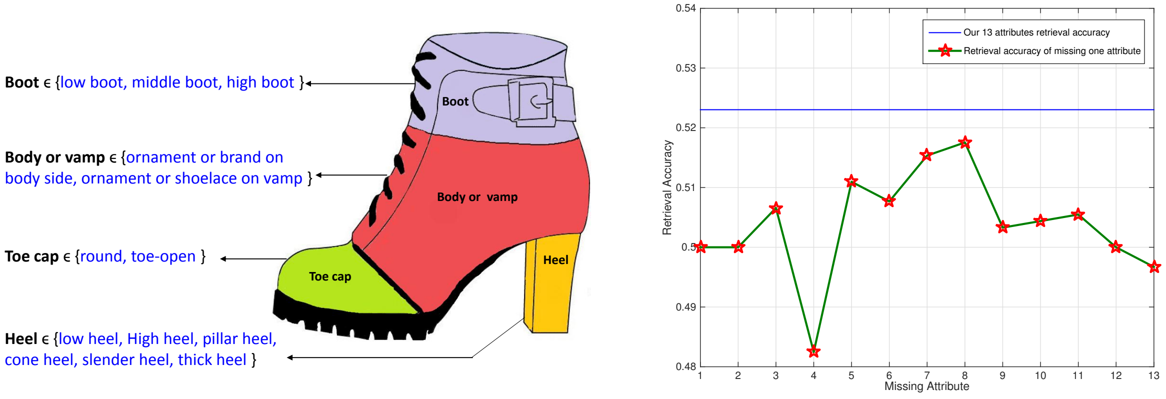
Content-based Image Retrieval, a problem that has been long studied by the computer vision community (see an excellent survey in [28]). Despite empowering various query and interaction modes, the main research focus remains to stay on the text-based queries. However, it is cumbersome to textually describe visual appearance such as complex object shape, and moreover it is imprecise due to demographic variation in descriptions. Instead, a simple free-hand sketch can speak for a "hundred" words without any language ambiguity and provides a far more expressive means of image search. Early approaches for sketch-based image retrieval (SBIR) mainly focused on feature engineering. Despite some success [9, 15], all assume pixel-level matching making them highly sensitive to alignment (and in turn work only with relatively accurate sketches). [16] conducted comparative and comprehensive experiments by evaluating traditional low-level feature descriptors (e.g., SIFT, HOG, SSIM, Shape Context, etc.) performance on SBIR, which demonstrated the cross-domain limitations of hand-crafted state-of-the-art image-based descriptors.

In order to address scalability, Cao *et al* [4] propose an edgel (edge pixel) structure to organize all database images. Their approach heavily relies on an edgel dictionary for the whole database, where each entry is represented by an edgel and several orientations. They measure sketch-image pair similarity by indexable oriented chamfer matching (IOCM), which makes it vulnerable to scale or orientation variance. Zhou *et al* [36] try to find the most salient part of an image in order to localize the correct region under cluttered background and do retrieval of a probe sketch based on this. However, determining saliency is a very hard problem and the accuracy of even the state-of-the-art saliency methods in natural images is low [22]), thus liming its reliability in practice.

To our knowledge, the only work that specifically tailored for *fine-grained* SBIR is [21], which scores sketch-image pair similarity in terms of four pose variations: viewpoint, zoom, configuration and body feature. Then a DPM [11] is employed as the representation to encode pose and coarse appearance in each domain, followed by a graph matching strategy for cross-domain pose correspondence. However, their criteria is pose rather than object detail-centric, and they lack a fine-grained SBIR dataset to validate on, so the efficacy for real fine-grained SBIR is unclear.

2.2. From Retrieval to Fine-Grained Retrieval

There have been extensive literature [13, 14, 29, 33] on category-level image retrieval, where they mostly em-



(a) (b)

Figure 2. (a) diagram of the proposed taxonomy of 13 part-aware attributes; different to conventional attributes defined at image-level, ours are spatially clustered within four semantic parts of a shoe, (b) per-attribute retrieval result, where a leave-one-out strategy is implemented; it shows each attribute is discriminative in its own right.

ploy image similarity frameworks. One way to build image similarity models is to first extract features like SIFT [24] and HOG [6], and then learn the image similarity models on top of these features, however, the performance is largely limited by the representation power of hand-crafted features. Another major drawback of traditional image retrieval is their inability to do instance-level retrieval, which requires distinguishing subtle differences between images of the same category. Yu *et al.* [35] for the first time explores *fine-grained* visual comparisons by applying a local learning approach based on relative attributes [26], like "the suspect is taller than him", "the shoes I want to buy are like these but more masculine". Inspired by above, very recently, Wang *et al.* [34] proposed a deep ranking model that learns fine-grained image similarity directly from images via learning to rank with image triplets. Despite some early success the problem remains largely unsolved, especially how they can be extended to work cross-domain as for the case of SBIR.

2.3. Fine-grained Attributes

Describing objects by their attributes [3, 10, 20, 27] has gained tremendous research attention recently, while comparatively little attention has been dedicated to the detailed structure of objects, particularly from a semantic viewpoint. Attributes capture information beyond the standard phraseology of object categories, instances, and parts, where *fine-grained* attributes further describe object parts with more detail. To our knowledge, there are only a few single-category datasets with *fine-grained* attribute annotations, for example, datasets related to detailed descriptions of birds [31], aircraft [30], and clothes [5]. We push this envelope by proposing a new dataset of fine-grained shoe attributes, not only on images but sketches as well.

3. A Large Scale Fine-Grained SBIR Dataset

In this section, we describe the collection of our fine grained shoe SBIR dataset with 304 images and 912 free-hand human sketches. Each image has three sketches corresponding to various drawing styles. This dataset provides a solid basis for all our learning tasks. Inspired by [8], we propose following criteria for the free-hand sketches and its corresponding image pairs collected in our dataset:

Exhaustive The images in our dataset cover most subcategories of shoes commonly encountered in day life.

Discriminative The shoe itself is unique enough and provides enough visual cues to be differentiated from others.

Practical The sketches are drawn by non-experts using their fingers on a touch screen, which resembles the real-world situations when sketches are practically used.

3.1. Defining a taxonomy of fine-grained shoe attributes

Attribute Discovery To identify a thorough list of fine-grained attributes for shoes, we start by extracting some from previous research on shoe images. Berg *et al.* [2] report the eight most frequent words that people use to describe a shoe, namely "front platform", "sandal style round", "running shoe", "clogs", "high heel", "great", "feminine" and "appeal". Kovashka *et al.* [19] further augment the list with another 10 relative attributes. It's noteworthy that the attributes they report are not particularly fine-grained in terms of *locality* and *granularity*, when compared with part-based ones defined in [30] for the category of airplanes. Some are functionality descriptions (e.g., sporty) or pure aesthetics (e.g., shiny) which make them fit to a typical attribute categorization paradigm. However, they provide a starting point to enable us to collect a fine-

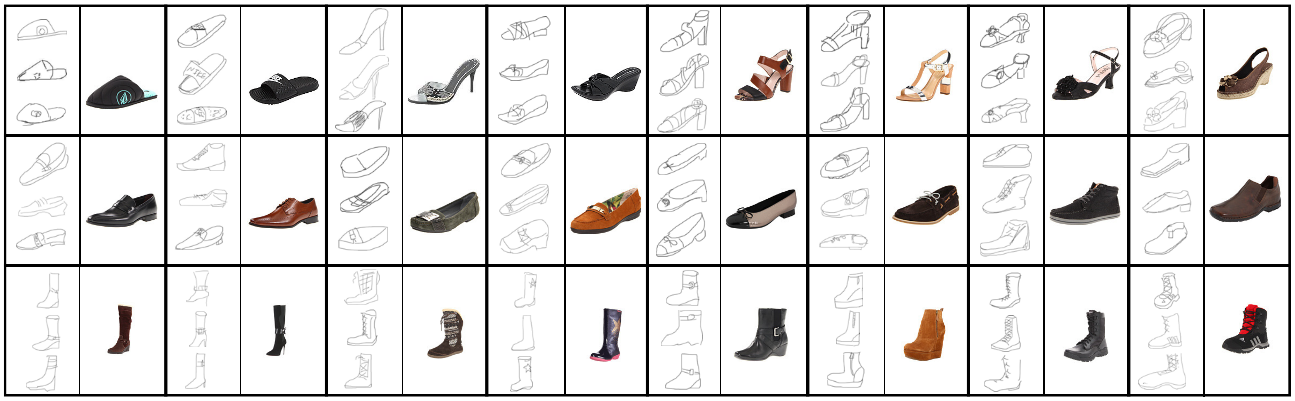


Figure 3. Representative sketch-photo pairs in our proposed *fine-grained* SBIR dataset, where each photo has three corresponding free-hand sketches drawn by different people. They highlight the abstract and iconic nature of sketches and differences in drawing ability among participants.

grained attribute inventory. We also mine the web (e.g., Amazon.com) and social media to find more key words and hashtags that people use to describe shoes, particularly those with higher degrees of *locality* and *granularity*. This gives us an initial pool of thirty fine-grained attributes.

Attribute Selection and Validation To determine which attributes are most suitable for our fine-grained SBIR task, we follow the "comparison principle" [30]. An attribute is considered informative only if it can be used to discriminate similar objects by pinpointing differences between them. This provides us two criteria for attribute selection (i) We omit shape or color-based attributes inappropriate to free-hand human sketches. (ii) We omit any attributes that jeopardize the overall retrieval accuracy when encoding both sketches and photos in terms of ground-truth attribute vectors. The selection criteria above leave us with 13 fine-grained shoe attributes, which we then cluster accordingly to one of the four parts of a shoe they are semantically attached to. Fig. 2 illustrates the selected attributes and their leave-one-out validation.

Collecting images The images are collected from the publicly available UT-Zap50K dataset [35] with 50,000 catalog shoe images from Zappos.com. From this, we choose a diverse set of 304 shoes from across all the subcategories, paying attention to include multiple inner detail variations.

Collecting sketches using crowdsourcing The main difficulties with collecting multiple sketches per image are: (i) ensuring sufficient diversity of sketching styles, and (ii) quality control on the sketches. To address this we use a crowdsourcing procedure, where each participant views an images, and draws the corresponding sketch including fine-grained object detail by recall. Multiple participants allow us to obtain different sketching styles for each image. Figure 3 illustrates some of the divergent drawing styles. Our sketches are distinguished from previous work by: (i) being finger-drawn on a tablet touch screen, which resembles a real-world application, and (ii) being in fine-grained correspondence to particular images.

Annotation With our finalized fine-grained SBIR dataset, we again use crowdsourcing to annotate both fine-grained attributes as well as parts which we will later use for strongly-supervised DPM training. To ensure high quality annotation and filter out bad workers, we randomly choose a certain proportion of the annotations for auditing, and participant's error rates directly determine their pay.

4. Methodology

Our learning approach is based on augmenting low- and mid-level feature representations with semantic attribute predictions that help distinguish subtle-but-important details [5, 7] in a domain invariant way (Sec. 4.1). This is then followed by enhancing these attributes to be part-aware (Sec. 4.2), and then integrating all three views of the image into a new robust representation (Sec. 4.3) for better fine-grained SBIR.

4.1. Feature and attribute extraction

Low-level feature extraction Histogram of Oriented Gradients (HOG) is extracted from shoes in both image and sketch domain. HOG is a ubiquitous descriptor that describes gradient information in a local patch. We extract HOG in a dense grid, and use it as a low-level sketch/photo representation. HOG was previously shown to be the best general-purpose feature representation for sketch [21, 16].

Learning an high-level attribute detection classifier From our ontology of $j = 1 \ldots A$ semantic attributes (Sec. 3.1), each training sketch/photo $\mathbf{x}$ in the dataset D is paired with attribute annotation $\mathbf{a}$, $D = \{\mathbf{x}_i, \mathbf{a}_i\}_{i=1}^{N}$. For each domain, and for each attribute j we then train a classifier $a_j(\cdot)$ to predict the presence/absence of the attribute using a binary support vector machine (SVM). Given the trained classifiers for each attribute, the A dimensional attribute representation for an sketch or image x is represented by stacking them as $\mathbf{a}(\mathbf{x}) = [a_1(\mathbf{x}), \ldots, a_A(\mathbf{x})]$.

4.2. Part-aware fine-grained SBIR

Our part detection mechanism has two purposes: (i) to generate a graph-model to encode the geometry of a shoe, and (ii) to support part-aware attribute detection.

Strongly-supervised DPM (SS-DPM) Model Instead of using the traditional DPM [11] where objects are represented by a coarse root HOG filter and several latent higher resolution part filters, we adopt SS-DPM here [1]. SS-DPM uses strong part-level supervision to improve the initialisation of the latent-SVM model parts rather than automatic heuristics. At this stage, a mixture of components is learned for each domain, which we denoted as $L^s = \{M_i^c\}_{i=1}^U$ for images and $L^p = \{M_j^c\}_{j=1}^U$ for sketches. For each M_i^c and M_j^c, we adopt the same feature learning and fusion procedures in Sec. 4.1, but in a localized way. Unlike [21], who use DPM on cross-domain pose correspondence and similarity scoring via graph matching, here, we simply aim to derive the shoe parts (bounding boxes), within which we detect fine-grained attributes.

Mid-level shoe structure representation To construct a more abstract and modality invariant representation based on shoe structure, we first need to detect shoe landmarks, which can be located by a strongly-supervised deformable part-based model (Section 4.2). Then a bank of relative coordinates derived from fully-connected graph model are used to represent our shoe structure information. Specifically, given L localized shoe landmarks (centre of the bounding boxes), a total of $\mathbf{s}(\mathbf{x}) = \frac{L\times(L+1)}{2}$ relative coordinates could be calculated by pairwise L2 distances for structure representation. This normally captures the distance between key features located and provides a novel view to discriminate between shoes.

Part-Aware Attribute Detection Once individual parts have been detected, these can be used to improve attribute detection compared to the holistic procedure outlined in Sec 4.1. Specifically, each attribute is associated with a localized shoe part (Fig 2), thus only the features from within the window of the detected part are used to predict the presence of that attribute. This requires the attribute detector to use the relevant cue and not inadvertently learn to detect irrelevant but correlated features from other parts. In this way we achieve de-correlated attribute learning that generalizes better at testing time, and in turn more accurate attribute detection accuracy that improves consequent retrieval performance.

4.3. Generating a combined representation

Three-view CCA With the availability of low-level features, attributes and shoe parts, we introduce a three-view CCA formulation to learn a new space that integrates all of these cues. Let X_x be the $(M + N) \times d$ dimensional matrix stacking the low-level feature representations $\mathbf{x}$ for all images and sketches and X_a is a $(M + N) \times A$ dimensional matrix stacking the attribute representations $\mathbf{a}(\mathbf{x})$ for all images and sketches, and X_s be the $(M + N) \times s$ dimensional matrix stacking the structural relative coordinates $\mathbf{s}(\mathbf{x})$. Then we find the projection matrices W_x and W_a and W_s that produce a single embedding [12] of these three views:

$$\begin{gathered}\underset{W_1,W_2,W_3}{min} \sum_{i,j=1}^{3} \left\|X_i W_i - X_j W_j\right\|_F^2 \\ subject\ to \quad W_i^T \Sigma_{ii} W_i = I, \quad w_{ik}^T \Sigma_{ij} w_{jl} = 0, \\ i,j = 1, \quad \ldots, 3, i \neq j, \quad k,l = 1, \ldots, c, \quad k \neq l\end{gathered} \tag{1}$$

where Σ_{ij} is the covariance between X_i and X_j and w_{ik} is the kth column of W_i, and c is the dimensionality of the desired CCA subspace. To better understand this objective function, let us consider its three terms:

$$\begin{gathered}\underset{W_x,W_a,W_s}{min} \left\|X_x W_x - X_a W_a\right\|_F^2 + \\ \left\|X_x W_x - X_s W_s\right\|_F^2 + \left\|X_a W_a - X_s W_s\right\|_F^2\end{gathered} \tag{2}$$

The first term tries to align corresponding low-level features and attributes, and the remaining two terms try to align with our part-aware structures. We argue that this will prove to provide a more robust and discriminative cross-domain representation for our fine-grained SBIR learning task. To solve this problem, we use the efficient generalized eigenvalue method of [12].

Using representation for fine-grained SBIR: After obtaining the estimated attributes $\mathbf{a}(\mathbf{x})$ and geometry $\mathbf{s}(\mathbf{x})$, we project them into the embedding space: $\mathbf{x}W_x$, $\mathbf{a}(\mathbf{x})W_a$, $\mathbf{s}(\mathbf{x})W_s$. Then concatenating all views to give our final $3c$ dimensional representation: $R(\mathbf{x}) = [\mathbf{x}W_x, \mathbf{a}(\mathbf{x})W_a, \mathbf{s}(\mathbf{x})W_s]$. Once our new robust and domain invariant representation is obtained for both sketches and images, matching a sketch $\mathbf{x}^s$ against a image dataset $D = \{\mathbf{x}_i^p\}_{i=1}^N$ is performed by nearest neighbor with L2 distance,

$$i^* = \underset{i}{\text{argmin}} \left|R^s(\mathbf{x}^s) - R^p(\mathbf{x}_i^p)\right| \tag{3}$$

Why Part-aware? Our goal is to learn attribute classifiers that fire only when the corresponding semantic property is present. In particular, we want them to generalize well even

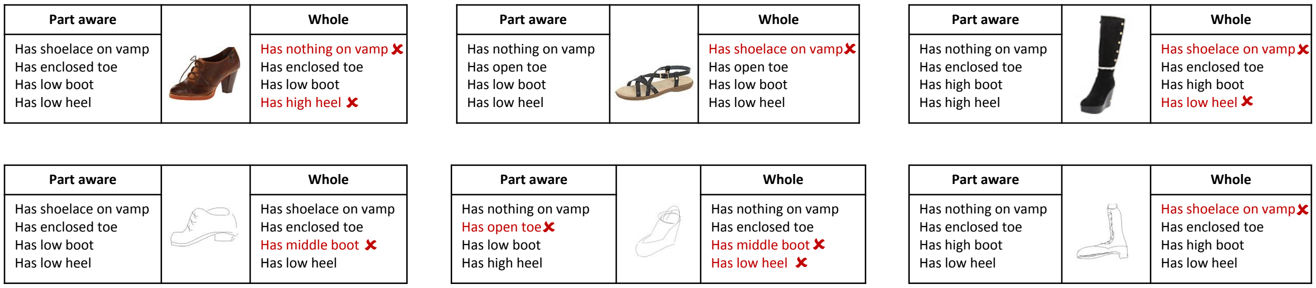


Figure 4. **Why part-aware?** We present some wrongly-detected attributes (red) when using whole image input, that have been corrected by our part-aware approach. (e.g., sometime, a tiny raise on the heel part does not necessary mean a low-heel shoe, instead, it may just an upward continuation of the sole part, which potentially makes it correlated with any attributes spatially in proximity of the shoe heels. Our part-aware approach can learn this subtlety semantically.)

when: (i) human free-hand sketches vary in shapes, scales, and width-height ratios. (ii) attribute co-occurrence patterns may differ from what is observed in training. The intrinsic pixel-oriented nature of SVM means it's prone to learn the wrong thing, even if it achieves high training accuracy, i.e., it may learn the properties that are correlated with the attribute of interest, rather than the attribute itself; and thus suffer if these correlations change at test time. In contrast, our part-aware model helps to achieve de-correlation and improve generalisation by detecting attributes on specific corresponding parts (details in Fig 4).

5. Experiments

5.1. Experimental settings

Preprocessing We first perform simple preprocessing to alleviate misalignment due to scale, aspect ratio, and centering. We downscale the height of the bounding boxes for both sketches and images to a fixed value of pixels while retaining their original aspect ratios. Then we locate the downscaled sketches and images to the centre of a $128 * 256$ blank canvas with rest padded by background pixels.

Low-Level Features For holistic low-level feature representation, we densely extract HOG on sketches and images on a $16 * 16$ grid, resulting in 4068 dimensional descriptor. We then use PCA to reduce HOG dimensions to 250. For part-level features required for fine-grained attribute detection, we constraint each part to be placed within a 128*64 patch, and pad by background pixels before performing the same feature extraction procedure as we do holistically.

SS-DPM training and detection Each SS-DPM is set to 4 mixture components and 4 parts per component, which in turn will deliver six relative coordinates for our shoe structural information. Unlike [1], all shoes in our dataset all share a uniform pose without partial occlusions. During detection, we choose the SS-DPM detection with the largest probability in each image and sketch. We use publicly available packages from [1] for full implementation with minor modifications. In Figure 6, we provide illustrations of part detection results on a few sketches and images in our dataset.

Training part-aware attribute detectors Using the 13 attribute taxonomy defined in Section 3.1, and the training procedure in Section 4.2, we produce a 13 dimensional binary attribute vector for each photo and sketch in the dataset.

Baselines (attribute detection): We compare against performance of the conventional means of using holistic features (**Whole-Image**), and using ground-truth part attributes as input (**Ground-Truth Part**). To further prove that our part-aware method somehow decorrelates the attributes, we evaluate against the state-of-the-art attribute decorrelation method introduced in [18], where they use semantic groups to encourage in-group feature sharing and between-group competition for features through a lasso multi-task learning framework. We compare with two variants of their method (i) similar to [18], when holistic image-wide features divided into 6 regular grids are used (**Weakly-Supervised (WS)-Decor**), and (ii) when ground-truth part annotations are supplied to extract part-level features (**Strongly-Supervised (SS)-Decor**). We also compare performance of strongly-supervised DPM against the original weakly-supervised DPM [11] which works without strong part annotations at training (**Weakly-Supervised (WS)-DPM**).

Baselines (fine-grained SBIR): We evaluate against when singular feature representations are used: (i) **Part-HOG**, where part-level HOG is employed, (ii) **Part-Attribute**, where only automatically detected part-aware attributes are utilized, and (iii) **Part-Structure**, where geometric part structure alone is used to retrieve. We also adopt the state-of-the-art two-view CCA method previously utilized to match facial sketches and caricatures to mugshot photo [25]. We compare with three pair-wise configurations to accommodate their two-view setting: **Part-HOG+Part-Attribute+2View-CCA**, **Part-HOG+Part-Structure+2View-CCA** and

Attribute	**Whole-Image**	**WS-Decor** [18]	**WS-DPM**	**SS-Decor** [18]	**Ours**	**Ground-truth part**
Round	90.33%	88.93%	90.96%	92.08%	**93.92%**	94.25%
Toe-open	90.33%	88.93%	90.96%	92.08%	**93.92%**	94.25%
Ornament or brand on body	65.45%	61.13%	66.39%	67.47%	**70.32%**	73.85%
Shoelace or ornament on vamp	63.03%	65.38%	64.10%	64.87%	**65.98%**	70.89%
Low heel	73.72%	70.74%	74.89%	73.11%	**75.44%**	77.25%
High heel	71.19%	77.60%	72.21%	**78.90%**	73.70%	76.72%
Pillar heel	82.64%	70.91%	82.50%	72.08%	**85.13%**	88.44%
Cone heel	63.71%	69.46%	64.11%	**74.85%**	67.53%	74.64%
Slender heel	82.76%	85.29%	84.02%	**88.24%**	86.54%	89.63%
Thick heel	88.24%	76.34%	88.89%	79.97%	**91.38%**	92.83%
Low boot	96.67%	90.94%	95.42%	95.82%	**97.08%**	98.04%
Middle boot	94.39%	87.91%	92.26%	91.67%	**95.78%**	96.92%
High boot	89.10%	88.98%	86.89%	91.41%	**91.15%**	93.23%
Average	80.89%	78.66%	81.05%	81.72%	**83.68%**	86.19%

Attribute	**Whole-Sketch**	**WS-Decor** [18]	**WS-DPM**	**SS-Decor** [18]	**Ours**	**Ground-truth part**
Round	80.80%	78.93%	80.14%	80.30%	**81.22%**	81.96%
Toe-open	80.80%	78.93%	80.14%	80.30%	**81.22%**	81.96%
Ornament or brand on body	54.91%	53.31%	56.81%	52.95%	**60.12%**	62.34%
Shoelace or ornament on vamp	73.02%	66.90%	**74.45%**	70.96%	72.99%	73.89%
Low heel	**66.45%**	63.20%	64.89%	64.21%	66.15%	74.29%
High heel	80.46%	79.86%	79.55%	**81.24%**	75.68%	83.29%
Pillar heel	69.86%	70.91%	67.89%	72.07%	**76.00%**	77.10%
Cone heel	59.79%	60.62%	60.12%	**64.07%**	63.10%	71.66%
Slender heel	78.51%	85.95%	76.87%	**87.38%**	79.71%	88.53%
Thick heel	69.93%	71.79%	65.21%	**74.73%**	70.60%	78.83%
Low boot	**92.51%**	87.49%	87.45%	87.70%	90.87%	94.04%
Middle boot	78.11%	77.74%	72.48%	79.65%	**84.03%**	85.51%
High boot	88.65%	86.32%	84.51%	**88.98%**	84.94%	90.32%
Average	74.91%	74.00%	73.12%	75.73%	**75.89%**	80.29%

Table 1. **Attribute detection using our part-aware method and other previous state-of-the-art method** On both image and sketch domains our method generally performs best, where some attributes actually outperform SS-Decor. One exception is that on sketch heel part, SS-Decor outperforms ours. Note however that SS-Decor required strong part annotation at testing time, whereas our SS-DPMs once trained works without part annotation at testing.

Part-Attribute+Part-Structure+2View-CCA.

5.2. Attribute detection

In this section, we evaluate our attribute-detection performance on both domains. In Table 1, we offer attribute detection accuracy (ten times random three-fold) for each of our sketch/image datasets. Overall, although many attributes are quite subtle, the average accuracies in the range 74%-84% clearly demonstrate that many of them can be reasonably reliably detected. More specifically, we can see that (i) all part-aware methods outperform whole-image, with ground-truth attributes offering the best performance, this further justifies the positive contribution of part localization, (ii) our method outperforms the state-of-the-art decorrelation method [18] on image and slightly on the more challenging sketch domain; it is however noteworthy that [18] required strong part annotations at testing, and our strongly-supervised DPM approach only utilized part annotation during training, and (iv) strongly-supervised DPM performs better than weakly-supervised alternative, again highlighting the importance of accurate part localization.

5.3. Fine-grained SBIR performance evaluation

We next perform quantitative evaluation on fine-grained SBIR. Given a probe sketch, we retrieve K images, and define a successful retrieval if there is a correct match within those K images. The results are illustrated by CMC curve in Fig 5, where we achieve an average of 52% @ $K = 10$, significantly outperforming traditional hand-crafted low-level features. In particular, we can observe that (i) singular feature representations are clustered at the bottom, with the relatively sparse structural feature being the worst, (ii) part-attribute alone is the best of all singular features with an accuracy of 35.33% @ $K = 10$, which surprisingly outperforms the part-hog+part-attribute CCA feature, and (iii) our three-view CCA method offers the best performance of all, with a more than 10% gain over the best two-view CCA method, and the closest to ground-truth retrieval using human attribute annotations (65.67% @ $K = 10$). In Fig 6, we present qualitative evidence that our part-aware fine-grained SBIR method can capture subtle variations across domains and deliver satisfying performance, e.g., in row 5 our method achieves more relevant images than the whole image approach by correctly matching the fine-grained details such as open vs closed heel, or high-heel vs platform.

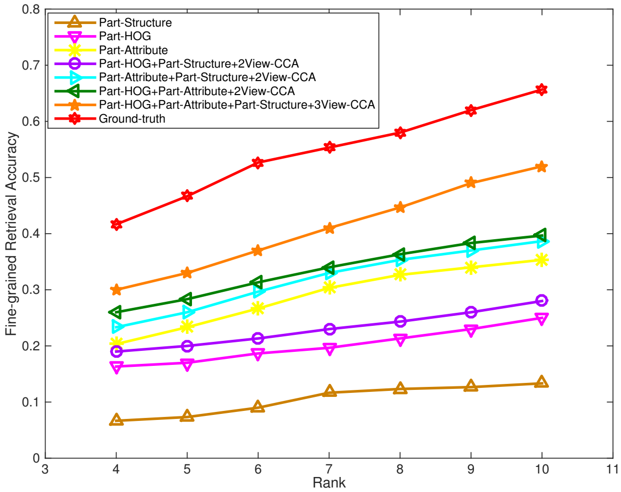


Figure 5. CMC curves for the proposed *fine-grained* SBIR framework, and comparisons with other baselines.

5.4. Analysis on different drawing styles

As shown in Figure 3, different sketches completed by different sketchers in our dataset have various levels of abstraction and deformation, or even different expressive interpretation on image-correspondence details. Thus, in this section, we present a pilot study on how diverse drawing styles could eventually affect our fine-grained SBIR outcome. More specifically, at dataset generation, we divided

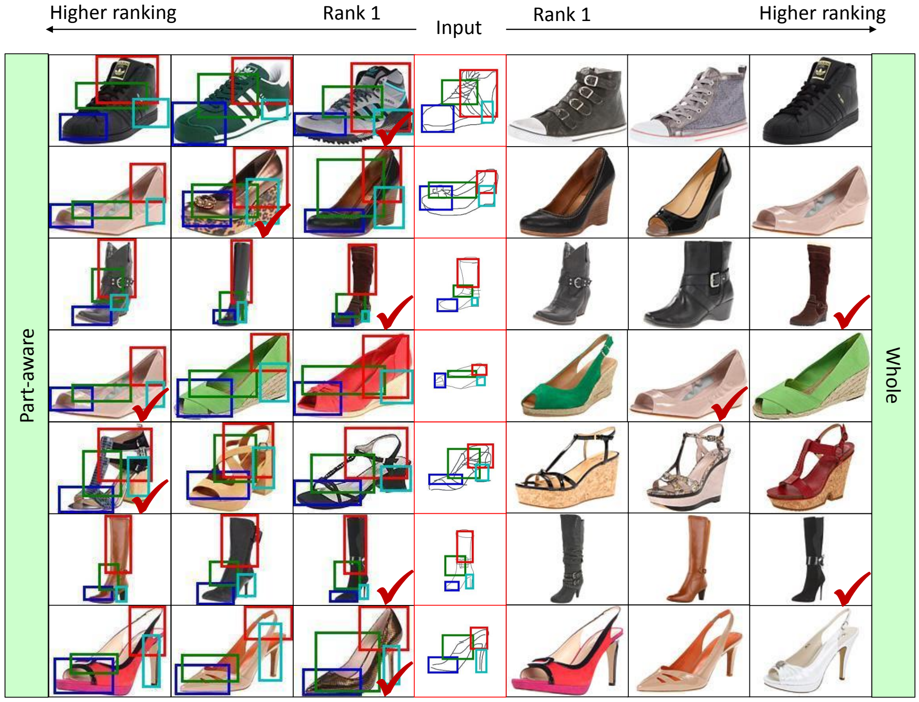


Figure 6. **An illustration of fine-grained SBIR result with and without our proposed part-aware method** Examples of some top ranking retrieval results given a probe sketch. Our part-aware method delivers sensible results, discriminating the *fine-grained* variations on a instance-level. Red-tick indicates ground-truth matching photo of the input sketch, which should be ranked as highly as possible. Part detection results of Strongly-Supervised DPM are shown using colour-coded bounding boxes.

Group	**Drawer 1**	**Drawer 2**	**Drawer 3**
No. 1	80%	70%	67%
No. 2	69%	74%	80%
No. 3	62%	54%	74%
No. 4	73%	65%	73%
No. 5	71%	79%	63%
No. 6	70%	75%	72%

Table 2. **Fine-grained SBIR results given different drawing styles.** Drawing style can affect the retrieval results significantly. This proves that our learning task is challenging due to the unrestricted non-expert free-hand sketches.

our participants into six groups, where each group is made up of three individuals. Each group was given the same set of images and draw a sketch for each images. Then some other participants manually annotate the fine-grained attributes that are present in each sketch and image. We examine and explore the sketching style of different people within each group through attribute-level SBIR, where the higher the sketch quality, the better the retrieval result. As can be seen in Table 2, the performance of fine-grained SBIR can vary dramatically due to different drawing styles across individuals. This result further highlights the challenging nature of the dataset and motivates future work to be carried out.

6. Conclusion

We investigated the practical problem of fine-grained sketch-based image retrieval (SBIR). For the first time, we studied the role of part-aware attributes. In doing so, we release a new SBIR dataset of shoes, where the dataset acquisition procedure was designed to closely resemble the realistic application scenario – users sketching with their fingers on a tablet some time delay after seeing a shoe. In particular, we proposed to detect attributes at part-level to construct a fine-grained semantic feature representation that not only works independently of visual domain, but also tackles the abstract and iconic nature of sketches. We further developed a three-view CCA space that captures low-level, middle-level and high-level information in a synergistic fashion. We demonstrated via extensive experiments that our strongly-supervised attribute detection framework can localize more accurate object parts hence outperforms state-of-the-art alternatives. Fine-grained SBIR results on the proposed dataset verified the effectiveness of part-aware attributes, both when used alone and synergistically with other features. In the future, we will investigate further the effect of style in fine-grained SBIR and study means of weakly-supervised part-level attribute detection and decorrelation.

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, pages 836–849. 2012.

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, pages 663–676. 2010.

[3] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, pages 438–451. 2010.

[4] Y. Cao, C. Wang, L. Zhang, and L. Zhang. Edgel index for large-scale sketch-based image search. In *CVPR*, pages 761–768, 2011.

[5] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, pages 5315–5324, 2015.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[7] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.

[8] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM TOG (Proceedings of SIGGRAPH)*, 2012.

[9] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. A descriptor for large scale image retrieval based on sketched feature lines. In *SBIM*, pages 29–36, 2009.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, pages 1627–1645, 2010.

[12] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, pages 210–233, 2014.

[13] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[14] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006.

[15] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, pages 1025–1028, 2010.

[16] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, pages 790–806, 2013.

[17] R. Hu, T. Wang, and J. Collomosse. A bag-of-regions approach to sketch-based image retrieval. In *ICIP*, pages 3661–3664, 2011.

[18] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, pages 1629–1636, 2014.

[19] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, pages 2973–2980, 2012.

[20] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[21] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[22] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.

[23] Y.-L. Lin, C.-Y. Huang, H.-J. Wang, and W.-C. Hsu. 3d subquery expansion for improving sketch-based multi-view image retrieval. In *ICCV*, pages 3495–3502, 2013.

[24] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[25] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li. Cross-modal face matching: Beyond viewed sketches. In *ACCV*, pages 210–225. 2014.

[26] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510, 2011.

[27] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 512–519, 2014.

[28] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1349–1380, 2000.

[29] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus. Learning invariance through imitation. In *CVPR*, pages 2729–2736, 2011.

[30] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, pages 3622–3629, 2014.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[32] C. Wang, Z. Li, and L. Zhang. Mindfinder: image search by interactive sketching and tagging. In *International Conference on World Wide Web*, pages 1309–1312, 2010.

[33] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *ICCV*, pages 428–435, 2009.

[34] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, pages 1386–1393, 2014.

[35] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, pages 192–199, 2014.

[36] R. Zhou, L. Chen, and L. Zhang. Sketch-based image retrieval on a large scale database. In *ICMR*, pages 973–976, 2012.