Synergistic Instance-Level Subspace Alignment for Fine-Grained Sketch-Based Image Retrieval

Citation for published version:
Li, K, Pang, K, Song, YZ, Hospedales, TM, Xiang, T & Zhang, H 2017, 'Synergistic Instance-Level Subspace Alignment for Fine-Grained Sketch-Based Image Retrieval', *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5908 - 5921. https://doi.org/10.1109/TIP.2017.2745106

Digital Object Identifier (DOI):
10.1109/TIP.2017.2745106

Published In:
IEEE Transactions on Image Processing

Synergistic Instance-Level Subspace Alignment for Fine-Grained Sketch-Based Image Retrieval

Ke Li[1,2] Kaiyue Pang[2] Yi-Zhe Song[2] Timothy M. Hospedales[2] Tao Xiang[2] Honggang Zhang[1]
[1]Beijing University of Posts and Telecommunications
[2]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 touch-screens. Three factors contribute to the challenging nature of the problem: (i) free-hand sketches are inherently abstract and iconic, making visual comparisons with photos 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 these challenges, we propose to bridge the image-sketch gap both at the high-level via parts and attributes, as well as at the low-level, via introducing a new domain alignment method. More specifically, (i) we contribute a dataset with 304 photos and 912 sketches, where each sketch and image is annotated with its semantic parts and associated part-level attributes. 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 provide pose-aligned sketch-image comparisons. To reduce the sketch-image gap when comparing low-level features, we also (iii) propose a novel method for *instance-level* domain-alignment, that exploits both subspace and instance-level cues to better align the domains. Finally (iv) these are combined in a matching framework integrating aligned low-level features, mid-level geometric structure and high-level semantic attributes. Extensive experiments conducted on our new dataset demonstrate effectiveness of the proposed method.**

***Index Terms*—Sketch-based Image Retrieval, Instance-level, Subspace alignment, Fine-grained, Cross-modal, Dataset.**

I. 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 [1], sketch-based image retrieval (SBIR) now has tremendous application potential.

Traditional computer vision methods for SBIR mainly focus on category-level retrieval, where intra-category variations are neglected. This is not ideal, since if given a specific shoe sketch (e.g., high-heel, toe-open) as query, it can return *any* shoe, including those with different part semantics (e.g., a flat running shoe). Thus *fine-grained* sketch-based image retrieval (FG-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 any practical commercial adoption of SBIR technology. For example, when one spots someone wearing a pair of shoes they really like but could not take a photo, instead of typing in textual descriptions which are both tedious and ambiguous, they could sketch their mental recollection of that shoe instead – FG-SBIR will find them the best matched pair of shoes, yet SBIR will return any shoe (which essentially renders sketching unnecessary since typing in the keyword 'shoe' into any text-based image retrieval engine will suffice). Figure 1 contrasts our *fine-grained* SBIR with traditional category-level SBIR systems.

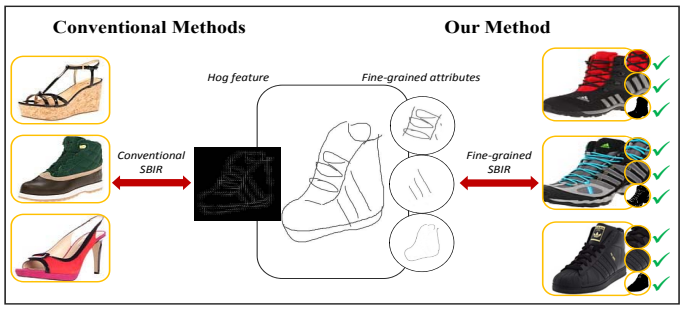


Fig. 1. Conventional SBIR operates at *category*-level, but *fine-grained* SBIR considers subtle details to provide instance-level retrieval. We propose a part-aware learning approach to train our semi-semantic representations based on a new large-scale *fine-grained* SBIR dataset of shoes. (Best viewed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

Fine-grained SBIR is challenging due to: (i) Free-hand sketches[1] 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. Above all, there is no 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 [2], [3], [4], [5], [6], [7] on retrieving images or 3d models based on sketches, typically with Bag Of Words (BOW) descriptors or advances

[1]A free-hand sketch is drawn without a refrence object or photo of the object present during drawing. The sketcher has to rely on either a mental recollection of the object seen before, or just the name of the object category. In the context of FG-SBIR, we focus on the former, i.e., without reference at hand but with recollection in memory.

thereof. Although BOW approaches are effective and scalable, they are weak at distinguishing *fine-grained* variations as they do not represent any semantic information, and suffer from sketch-image domain shift. **Semantics:** Recently, approaches to *fine-grained* SBIR have included Deformable Parts Model (DPM)-based part modeling in order to retrieve objects in specific poses [8]. However, for practical SBIR in commercial applications, we are more interested in distinguishing subtly different object sub-categories or attributes rather than poses. *Attributes* have recently been used to help drive *fine-grained* image retrieval by identifying subtle yet semantic properties of images [9], [10]. Moreover, such attributes may provide a route to bridge the sketch/image semantic gap, as they are domain invariant if reliably detected (e.g., a high-heel shoe is 'high-heel' regardless if depicted in an image or sketch). However, they suffer from being hard to predict due to spurious correlations [11]. **Domain Shift:** Low-level feature encodings, whether conventional Histogram of Oriented Gradients (HOG)-BOW or deep features suffer from domain shift across the sketch-image domains, thus significantly reducing matching accuracy. Aiming to address this, there is extensive work on domain-adaptation such as subspace alignment [12], [13] – typically used to transfer a within-domain classifier to another domain; and cross-domain projections [14], [15], [16], [17], [18], [19], [20] – typically used for cross-domain matching. **Instance-level Alignment:** Yet all prior work on domain alignment operate on category-level, consequently making them not directly applicable to the *fine-grained* instance-level retrieval task in hand – *fine-grained* SBIR. In this work, we propose an instance-level domain alignment method that is specifically designed for *fine-grained* SBIR. More specifically, we reduce the instance-level sketch-image gap by combining the favorable properties of subspace based domain-adaptation and instance-based projection methods. Finally, we bring the *fine-grained* feature alignment and high-level semantic matching strategies together to provide effective FG-SBIR.

To address the domain gap at a high semantic level we work with parts and attributes. We 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. We then propose a part-aware SBIR framework that addresses the *fine-grained* SBIR challenge by identifying discriminative attributes and parts, and then building a 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). To address the domain gap at the level of low-level features, we propose a novel *fine-grained* domain alignment method that searches for an alignment that both (i) robustly aligns the domains' subspaces to make them directly comparable, as well as (ii) provides *fine-grained* instance level alignment across the domains. At retrieval time, based on these strategies to align the domains at both high and low-levels, we can simply apply nearest neighbour matching to to retrieve images most similar to the probe sketch. We demonstrate the superiority of our framework on FG-SBIR through in-depth comprehensive and comparative experiments.

The contributions of our work are:

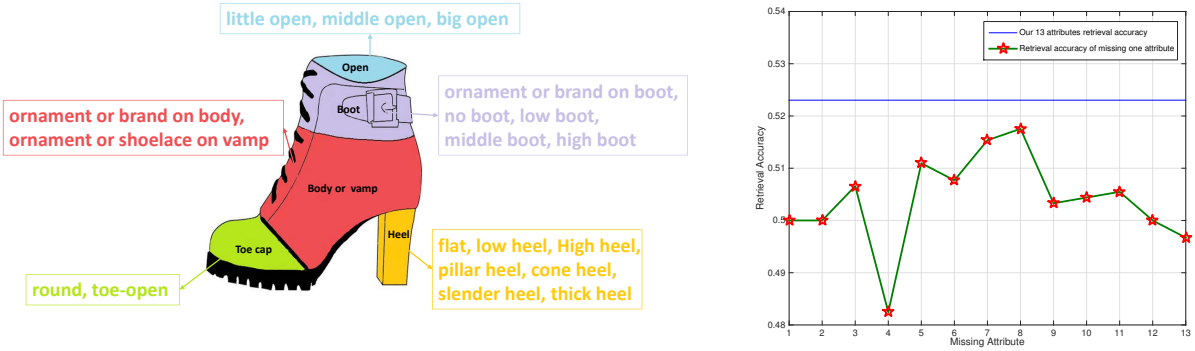
- We contribute a FG-SBIR shoe dataset with free-hand human sketches and photos, as well as *fine-grained* attribute annotations.
- We propose a part-aware paradigm that allows FG-SBIR attribute detection.
- A novel instance-level cross-modal domain alignment method is developed to robustly and accurately bridge the domain gap by requiring both subspace and instance-level alignment.
- Bringing these components together, we demonstrate that exploiting representation at both low-and high levels provides significantly improved FG-SBIR performance.

II. Related Work

1) Sketch-based Image Retrieval: Text-based queries can be efficient by using keyword tags to indicate the presence of salient objects or concepts. However, it can become cumbersome when describing visual appearance such as complex object shape or style and imprecise due to wide demographic variations. Instead, a simple free-hand sketch can speak for a "hundred" words without any language ambiguity and provide a far more expressive means of image search. Despite some success [5], all assume pixel-level matching, making them highly sensitive to alignment (and in turn work only with relatively accurate sketches). [4] conducted comparative and comprehensive experiments by evaluating traditional low-level feature descriptors (e.g., SIFT, HOG, 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* [3] 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, which makes it vulnerable to scale or orientation variance. Zhou *et al* [21] 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.

Existing work tailored for *fine-grained* SBIR is quite limited [8], [23]. [8] uses a DPM to represent objects in sketch and image domain, followed by graph-matching to establish correspondence. However, this is designed for matching object pose rather than *fine-grained* object details. [23] employs a multi-branch deep neural network to learn a representation that bridges sketch-image gap, at the instance level. Although very successful, the scalability is limited in practice by the need to manually annotate $O(N^3)$ triplets, and the computational requirements of training them; we show that our *fine-grained* domain alignment method performs comparably or better to [23] while having much more reasonable annotation and computational requirements.



(a) (b)

Fig. 2. (a) Proposed taxonomy of 13 part-aware attributes; different to conventional attributes defined at image-level, ours are localized within four semantic parts of a shoe, (b) Per-attribute retrieval result, where a leave-one-out strategy is implemented; it shows each attribute contributes to shoe discrimination.

2) From Retrieval to Fined-grained Retrieval: There has been extensive research [24], [25], [26], [27] on category-level image retrieval. A common approach is to first extract features like SIFT and HOG, and then learn image similarity models on top of these. However, the performance is largely limited by the representation power of hand-crafted features. And importantly, this approach is not effective for *fine-grained* retrieval, which requires distinguishing subtle differences between images within the same category. Yu *et al*. [10] for the first time explore *fine-grained* visual comparisons by applying a local learning approach based on relative attributes [28], like "the suspect is taller than him", "the shoes I want to buy are like these but more masculine". Inspired by this, Wang *et al*. [29] 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 in terms of how they can be extended to work cross-domain as for the case of SBIR.

3) Fined-grained Attributes: Describing objects by their attributes [30], [31], [32], [33] 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 [34], aircraft [35], and clothes [36]. We push this envelope by proposing a new dataset of *fine-grained* shoe attributes, not only on images but sketches as well.

4) Cross-Modal Alignment: Cross-modal alignment has drawn increasing attention due to the growing prevalence of multi-modal data. Three types of approaches can be identified according to which type of supervision they use: instance-level, category-level, or unsupervised. *Instance-level*: Canonical Correlation Analysis (CCA) [14], Partial Least Square (PLS) [16] and Bilinear Model (BLM) [37] are popular approaches that aim to map corresponding images from different modalities (e.g., sketch and photo) to a common subspace where corresponding instances are highly correlated. *Category-level*: Sharma et. al. [19] proposed Generalized Multi-view Linear Discriminant Analysis (GMLDA) and Generalized Multi-view Marginal Fisher Analysis (GMMFA) as the multi-view counterparts of Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA), respectively. Learning Coupled Feature Spaces for Cross-modal Matching (LCFS) [20] learns a low-rank projection to select relevant features for projecting across domains. These methods additionally use category-level supervision. For example, to say that in the learned space, same-category images should be near and different-category images should be far. *Unsupervised*: Unsupervised methods such as Domain Adaptation Subspace Alignment (DA-SA) [13] and Transfer Joint Matching (TJM) [38] aim to align domains without using class labels or instance pairs via subspace or maximum mean discrepancy (MMD)-based alignment. Recently, Xu et. al. [39] performed a comparative study of different cross-modal alignment methods on the FG-SBIR task, and found LCFS and CCA to be the most effective existing methods.

5) Cross-Modal Alignment for Fine-Grained Retrieval: For FG-SBIR we are interested in intra-category retrieval: finding *that* specific shoe that you saw, not finding shoes instead of chairs. Thus exploiting category-level supervision in alignment is less relevant. Instance-level and unsupervised methods can be applied. But the former misses the holistic cue from the whole dataset distribution, and the latter misses the fine-grained detail of instance-level correspondence. Our contribution is therefore to propose a 'fine-grained subspace alignment' (FG-SA) method that exploits *both* the holistic dataset-level alignment intuition used by methods such as DA-SA along with the specific instance-level matching used by methods such as CCA and PLS. However, unlike typical instance-level methods like CCA/PLS which only require that corresponding instances are highly correlated, we explore using instance level cues discriminatively: to also require that mismatching instances should be dissimiliar, i.e., similar to the class-separability intuition used by some category-level methods but applied to individual instance correspondences.

III. A 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 sketch correspondings to various drawing styles. Inspired by [1], we propose the following

criteria for the free-hand sketches and their corresponding photo images 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.

A. Defining a Taxonomy of Fine-grained Shoe Attributes

Attribute Discovery To identify a comprehensive list of *fine-grained* attributes for shoes, we start by extracting some from previous research on shoe images. Berg *et al*. [40] 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*. [41] further augment the list with another 10 relative attributes. It is 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 [35] for the category of airplanes. Some are functional (e.g., sporty) or aesthetic (e.g., shiny) descriptions which make them fit to a typical attribute categorization paradigm. However, they provide a starting point to enable us to collect a *fine-grained* attribute inventory. We also mine the web (e.g., Amazon.com) and social media to find more key words and hash tags 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" [35]. 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. Figure 2 illustrates the selected attributes and their leave-one-out validation.

Collecting Images The images are collected from the publicly available UT-Zap50K dataset [10] with 50,000 catalogue shoe images from Zappos.com. From this, we choose a diverse set of 304 shoes from across all the subcategories, paying attention to including 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 image, 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 the diversity of sketch styles obtained while being in *fine-grained* correspondence to a given image. To control the quality, each image was drawn by multiple workers and for each image the top three best-drawn sketches were kept.

Annotation To annotate our final *fine-grained* SBIR dataset, we again use crowdsourcing for both *fine-grained* attributes as well as parts which we will later use for strongly-supervised DPM training.

IV. 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 [36], [9] on a domain invariant way (Sec. IV-A). This is then followed by enhancing these attributes to be part-aware (Sec. IV-B), and then aligning the low-level feature views of both sketch and image domains via both their instances and subspaces (Sec. IV-C), and finally combining all three views of the image to achieve robust and accurate matching (Sec. IV-D) for better *fine-grained* SBIR.

A. 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/image representation. HOG was previously shown to be the best general-purpose feature representation for sketch [8], [4].

Learning a High-level Attribute Detector Each training sketch/image in our dataset is paired with attribute annotation. For each domain, and for each attribute j we can train a classifier $a_j(.)$ to predict the presence/absence of the attribute given an image/sketch using a binary support vector machine (SVM). The final attribute representation is then produced by stacking classifier outputs into a single vector.

This strategy provides a domain invariant representation, but when applied to raw HOG descriptors, attribute detection is unreliable due to lack of alignment, and feature selection (most of the input image is irrelevant to any given local attribute). Thus we improve attribute detection by explicitly detecting object parts.

B. A Part-aware Approach

Our part detection mechanism serves 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 Deformable Parts Model (SS-DPM) Instead of using the traditional DPM [42] where objects are represented by a coarse root HOG filter and several latent higher resolution part filters, we adopt SS-DPM here [43]. SS-DPM uses strong part-level supervision to improve the initialisation of the latent-SVM model parts rather than automatic heuristics. Unlike [8], which uses DPM for cross-domain pose correspondence via graph matching, we only aim

Fig. 3. Representative sketch-image pairs in our proposed *fine-grained* SBIR dataset, where each image 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.

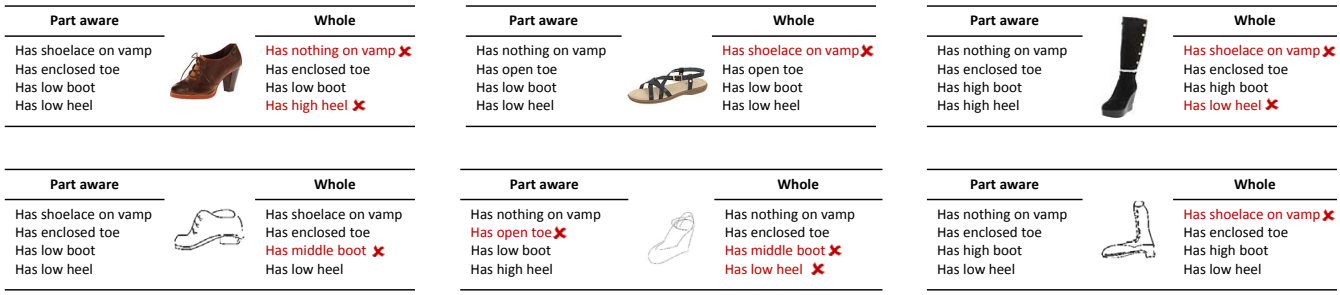


Fig. 4. **Why part-aware?** We present some wrongly-detected attributes (red) when using whole image input, that have been corrected by our part-aware approach. (e.g., sometime, a tiny raise on the heel part does not necessary mean a low-heel shoe, instead, it may just an upward continuation of the sole 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.)

to derive the appropriate shoe parts (bounding boxes), within which we detect corresponding *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 are located by the strongly-supervised-DPM model above. 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 $\frac{L\times(L+1)}{2}$ relative coordinates denoted $\mathbf{s}(\mathbf{x})$ are calculated by pairwise L2 to provide a mid-level structure representation encoding the geometry via distances between pairs of key features on the shoe.

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 IV-A. 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 [11]. 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.

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 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 applied on a global feature means that it is prone to learn the wrong thing, even if it achieves high training accuracy. E.g., 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).

Summary The method thus far provides a view of sketches and photos which is modality invariant by design due to encoding via accurately detected high-level attributes. Nevertheless, this high-level information alone is insufficient to discriminate all *fine-grained* sketch-image comparisons: either because of outstanding imperfections in attribute detection, or because multiple shoes share a given attribute vector encoding. Thus we next turn to cross-modal alignment of low-level features in order to complement our high-level representation above.

C. Fine-grained Subspace Alignment

In this section, we detail our domain-alignment method which aligns the low-level feature (HOG) views of sketch and image domains in terms of both subspaces and *fine-grained* instance pairs.

Subspace Generation We represent every sketch and image as a D-dimensional z-normalized vector (i.e. zero mean). Thus $\mathbf{f}_I$ and $\mathbf{f}_S \in \mathcal{R}^{N\times D}$ are the matrices stacking all N images in each modality, with $\mathbf{f}_I^i$ ($\mathbf{f}_S^i$) denoting the i-th row in $\mathbf{f}_I$ ($\mathbf{f}_S$). Using Principal Component Analysis (PCA), we then generate a subspace for each domain, represented by the d eigenvectors corresponding to the d largest eigenvalues. These subspaces are denoted by $\mathbf{X}_I$ and $\mathbf{X}_S$ ($\mathbf{X}_I, \mathbf{X}_S \in \mathcal{R}^{D\times d}$), and they are orthonormal ($\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S = \mathbf{I}_d$ and $\mathbf{X}_I^{\mathrm{T}}\mathbf{X}_I = \mathbf{I}_d$, where $\mathbf{I}_d$ is identity matrix of size d). We next align the subspaces $\mathbf{X}_I$ and $\mathbf{X}_S$ of the two domains.

Domain-level Subspace Alignment Fernando *et al.* [13] proposed to align subspaces $\mathbf{X}_I$ and $\mathbf{X}_S$ to make data in each domain ($\mathbf{f}_S^i$ and $\mathbf{f}_I^i$) more comparable. To align the subspaces, a $d \times d$ transformation matrix $\mathbf{M}$ is learned by minimizing the following Bregman matrix divergence:

$$F_1(\mathbf{M}) = \|\mathbf{X}_I\mathbf{M} - \mathbf{X}_S\|_F^2 \tag{1}$$

where $\|.\|_F^2$ is the Frobenius norm. The transformation learned in Eq. 1 robustly aligns the axes of variation of the two domains, making them more comparable. However it does not exploit available information of sketch-photo correspondences.

Instance-level Subspace Alignment In contrast to domain-adaptation approaches that often operate on *subspaces*, cross-modal matching problems such as *fine-grained* SBIR search for projections that bring individual *instances* into correspondence. This is because the often subtle visual and structural differences on fine-grained research tasks can not be effectively distinguished using domain-level subspaces. As a result, one needs to take *instance-level* feature into consideration. Let's denote $\mathbf{f}_I^i\mathbf{X}_I\mathbf{M}$, and $\mathbf{f}_S^i\mathbf{X}_S$ as the i-th image and sketch instance projection feature respectively, and writing as $\mathbf{f}_I^i\mathbf{X}_I\mathbf{M}-\mathbf{f}_S^i\mathbf{X}_S$ the difference of transformed i-th image-sketch pair. We additionally optimize the $l_{2,1}$ norm of pair distances:

$$F_2(\mathbf{M}) = \sum_{i=1}^{N} \left\|\mathbf{f}_I^i\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2. \tag{2}$$

Optimizing this $l_{2,1}$-norm is more robust to outlying instances than conventional F-norm [44], so if some badly drawn sketches are impossible to align with their corresponding image, then the detriment to the learned mapping is small.

Discriminative Instance-level Subspace Alignment The method introduced in Eq. 2 minimizes the distances between corresponding photo-sketch pairs, but better results may be achieved if we further prefer a discriminative alignment where the distances between matching pairs are smaller than those between mismatching photo-sketch pairs – an intuition widely exploited in metric learning [45], [46]. To exploit this stronger constraint, we optimise the difference between the average $l_{2,1}$ distances of corresponding and mismatching pairs:

$$\begin{aligned} F_2(\mathbf{M}) = & \sum_{i=1}^{N} \left\|\mathbf{f}_I^i\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2 \\ & -\lambda * \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{f}_I^j\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2 \end{aligned} \tag{3}$$

The above constraint requires that matching pairs are closer than mismatching pairs on average, but even if this constraint is met there could be many individual examples for which mismatching photos and sketches are closer than matching ones. A stronger version of the constraint is therefore to require that each individual positive pairing is closer than all alternative negative pairings:

$$\begin{aligned} F_2(\mathbf{M}) = \frac{1}{N} * \sum_{i=1}^{N}\sum_{j=1}^{N} [& \left\|\mathbf{f}_I^i\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2 \\ & - \left\|\mathbf{f}_I^j\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2]_+ \end{aligned} \tag{4}$$

where $[.]_+$ is hinge loss. Optimising the transformation $\mathbf{M}$ in this formula guarantees that for each sketch, the difference between matching image-sketch pair is less than any mismatching pairing. That is, when the loss $F_2(\mathbf{M})$ (Eq. 4) reaches zero, the training set has 100% matching accuracy. However, as this is a much harder constraint to meet compared to Eq. 3, it is possible that more overfitting occurs, so the testing performance is not necessarily better than that of Eq. 3.

Simultaneously Aligning Subspaces and Instances Our overall objective is to improve cross-domain matching by combining subspace alignment and instance alignment intuitions. To this end we explore three method variants corresponding to the three instance-alignment options proposed in Eqs. 2-4. These are to minimize the loss $F(\mathbf{M})$ in the following:

Fine-grained Subspaces Alignment Method 1:

$$F(\mathbf{M}) = \|\mathbf{X}_I\mathbf{M} - \mathbf{X}_S\|_F^2 + \lambda * \sum_{i=1}^{N} \left\|\mathbf{f}_I^i\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2 \tag{5}$$

Fine-grained Subspaces Alignment Method 2:

$$\begin{aligned} F(\mathbf{M}) = \|\mathbf{X}_I\mathbf{M} - \mathbf{X}_S\|_F^2 + \lambda * \textstyle\sum_{i=1}^{N} \\ (\left\|\mathbf{f}_I^i\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2 - \frac{1}{N}\textstyle\sum_{j=1}^{N} \left\|\mathbf{f}_I^j\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2) \end{aligned} \tag{6}$$

Fine-grained Subspaces Alignment Method 3:

$$\begin{aligned} F(\mathbf{M}) = \|\mathbf{X}_I\mathbf{M} - \mathbf{X}_S\|_F^2 + \lambda * (\tfrac{1}{N} * \textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} \\ [\left\|\mathbf{f}_I^i\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2 - \left\|\mathbf{f}_I^j\mathbf{X}_I\mathbf{M} - \mathbf{f}_S^i\mathbf{X}_S\right\|_2]_+) \end{aligned} \tag{7}$$

More specifically, Eq. 5 performs domain-level alignment while minimizing pair-wise distances between sketch-photo pairs after projection (Eq. 2); Eq. 6 additionally takes into account mismatching pairs by enforcing average distance between matching pairs to be smaller than those between mismatching pairs (Eq. 3); finally, Eq. 7 further constrains instance-level alignment by asking all positive pairs to be closer than all negative pairs (Eq. 4).

The objectives in Eqs. 5-7 can be optimized by stochastic gradient descent. But the $l_{2,1}$ distance means the losses have many non-differentiable points, requiring optimization by sub-gradient. The optimization procedure is summarized in Algorithm 1.

D. Matching Procedure

Given the high-level representation and low-level alignment learned in Sec. IV-A-IV-C, we can effectively match sketches and photos as follows. For a given image and sketch i and j we project their low-level features into their respective subspaces,

and align the images with the sketches, obtaining the new representations: $\mathbf{R}_I^i = \mathbf{f}_I^i \mathbf{X}_I \mathbf{M}$ and $\mathbf{R}_S^j = \mathbf{f}_S^j \mathbf{X}_S$. We also obtain part-based attribute $\mathbf{a}(\cdot)$ and geometry $\mathbf{s}(\cdot)$ information from Section IV-A and Section IV-B. Combining these three views, we quantify the distance between pair i and j as:

$$D(i,j) = \left\| \mathbf{R}_I^i - \mathbf{R}_S^j \right\|_2 + w_1 * \left\| \mathbf{a}(\mathbf{f}_I^i) - \mathbf{a}(\mathbf{f}_S^j) \right\|_1 + w_2 * \left\| \mathbf{s}(\mathbf{f}_I^i) - \mathbf{s}(\mathbf{f}_S^j) \right\|_2 \tag{8}$$

where w_1, w_2 are the weighting factors allocated for the 3 views, obtained by a greedy search.

V. Experiments

A. Experimental settings

In the first set of experiments, we focus on evaluating attribute detection accuracy, the significance of detecting part localised attributes, and compare overall matching performance using our basic *fine-grained* subspace alignment (FG-SA) Method 1 to a wide variety of prior cross-modal matching work.

Datasets We perform experiments on three *fine-grained* SBIR datasets:

- **912 shoe sketch database:** This is the database presented in this work and used in earlier experiments [47]. We perform 3-fold cross validation, and report average result over 10 random splits.
- **419 Shoes and 297 Chairs database [23]:** We use the same shoe and chair FG-SBIR datasets introduced by [23]. We follow their experiment setup and use 304 pairs of shoes and 200 chairs for training, with the rest for testing.
- **Sketchy [48]:** The most recent and largest *fine-grained* SBIR dataset consisting of 74,425 sketches and 12,500 gallery photos spanning 125 categories. In each category, there are 100 object instances, and each instance has one photo and 5 or more corresponding sketches. We use the same training dataset and testing dataset introduced by [48], with the same 90/10 train/test split.

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 place the downscaled sketches and images at 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 $D = 4608$ dimensional descriptor.

For part-level features required for *fine-grained* attribute detection, we constrain each part to be placed within a bounding box (size 64×64 patch for 'toe-cap', 'body', 'body' and 'heel' parts; and a 128×64 patch for 'boot' part) before performing the same 16×16 dense grid feature extraction as we do holistically. This results in $D = 576, 576, 576, 1152$ dimensional part descriptors respectively.

Input: $\mathbf{f}_I, \mathbf{f}_S \in R^{N*D}$
$\mathbf{X}_I, \mathbf{X}_S \in R^{D*d}$
step size: α
Output: $\mathbf{M} \in R^{d*d}$

Let $t = 1$,Initialize $\mathbf{M}_t = \mathbf{X}_I^{\mathrm{T}} \mathbf{X}_S$
while $F(M_t) - F(M_{t+1}) > 0.01$ **do**
  Calculate the block diagonal matrices $B_t \in N * N$,where the i-th diagonal block of B_t is $\frac{1}{2\sqrt[2]{\left\|\mathbf{f}_I^i \mathbf{X}_I \mathbf{M}_t - \mathbf{f}_S^i \mathbf{X}_S\right\|_2^2 + \zeta}} I_j$,where $\zeta \to 0$.
  Calculate a series of diagonal matrices $E^j \in N * N$, where $j \in [1, N]$.The i-th diagonal block of E^j is $\frac{1}{2\sqrt[2]{\left\|\mathbf{f}_I^j \mathbf{X}_I \mathbf{M}_t - \mathbf{f}_S^i \mathbf{X}_S\right\|_2^2 + \zeta}} I_j$.
  Calculate the diagonal matrix $Q \in N * N$, the j-th diagonal block of Q is $\left\|E^j\right\|_1$. Calculate matrix $P \in N * N$, the j-th row i-th column element of P is $\frac{1}{2\sqrt[2]{\left\|\mathbf{f}_I^j \mathbf{X}_I \mathbf{M} - \mathbf{f}_S^i \mathbf{X}_S\right\|_2^2 + \zeta}}$.
  For **Method 1**: $\frac{\partial F(\mathbf{M}_t)}{\partial \mathbf{M}} = (\mathbf{X}_I^{\mathrm{T}} \mathbf{X}_I + \lambda * (\mathbf{f}_I \mathbf{X}_I)^T B_t \mathbf{f}_I \mathbf{X}_I) M_t - (\mathbf{X}_I^{\mathrm{T}} \mathbf{X}_S + \lambda (\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} B_t \mathbf{f}_S \mathbf{X}_S)$.
  For **Method 2**: $\frac{\partial F(\mathbf{M}_t)}{\partial \mathbf{M}} = (\mathbf{X}_I^{\mathrm{T}} \mathbf{X}_I + \lambda * ((\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} B \mathbf{f}_I \mathbf{X}_I - \frac{1}{N} (\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} Q \mathbf{f}_I \mathbf{X}_I)) \mathbf{M}_t - (\mathbf{X}_I^{\mathrm{T}} \mathbf{X}_S + \lambda * ((\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} B \mathbf{f}_S \mathbf{X}_S - \frac{1}{N} (\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} P \mathbf{f}_S \mathbf{X}_S))$.
  For **Method 3**: For each $i \in [1, ..., N]$, calculate $h_i \in [h_{min}, ..., h_{max}]$, where $\left\|\mathbf{f}_I^i \mathbf{X}_I \mathbf{M} - \mathbf{f}_S^i \mathbf{X}_S\right\|_2 - \left\|\mathbf{f}_I^{h_i} \mathbf{X}_I \mathbf{M} - \mathbf{f}_S^i \mathbf{X}_S\right\|_2 > 0$. h_l is the length of $[h_{min}, ..., h_{max}]$
  Calculate a series of diagonal matrices $E_h^{h_i} \in N * N$, where $h_i \in [h_{min}, ..., h_{max}]$.The i-th diagonal block of $E_h^{h_i}$ is $\frac{1}{2\sqrt[2]{\left\|\mathbf{f}_I^{h_i} \mathbf{X}_I \mathbf{M} - \mathbf{f}_S^i \mathbf{X}_S\right\|_2^2 + \zeta}} I_{h_i}$.
  Calculate the diagonal matrix $Q_h \in h_l * h_l$, the i-th diagonal block of Q_h is $\left\|E_h^{h_i}\right\|_1$.
  Calculate the matrix $P_h \in h_l * N$, the j-th row i-th column element of P is $\frac{1}{2\sqrt[2]{\left\|\mathbf{f}_I^j \mathbf{X}_I \mathbf{M} - \mathbf{f}_S^i \mathbf{X}_S\right\|_2^2 + \zeta}}$,where $j \in [h_{min}, ..., h_{max}], i \in [1, N]$.
  $\mathbf{f}_I^h \in h_l * D$ is a subset of $\mathbf{f}_I$, the i-th row of $\mathbf{f}_I^h$ is the h_i row of $\mathbf{f}_I$,where $h_i \in [h_{min}, ..., h_{max}]$.
  $\frac{\partial F(\mathbf{M}_t)}{\partial \mathbf{M}} = (\mathbf{X}_I^{\mathrm{T}} \mathbf{X}_I + \lambda * \frac{1}{N} * (h_l (\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} B \mathbf{f}_I \mathbf{X}_I - (\mathbf{f}_I^h \mathbf{X}_I)^{\mathrm{T}} Q_h \mathbf{f}_I^h \mathbf{X}_I)) \mathbf{M}_t - (\mathbf{X}_I^{\mathrm{T}} \mathbf{X}_S + \lambda * \frac{1}{N} * (h_l (\mathbf{f}_I \mathbf{X}_I)^{\mathrm{T}} B \mathbf{f}_S \mathbf{X}_S - (\mathbf{f}_I^h \mathbf{X}_I)^{\mathrm{T}} P_h \mathbf{f}_S \mathbf{X}_S))$
  Calculate $\mathbf{M}_{t+1} = \mathbf{M}_t - \alpha \frac{\partial F(\mathbf{M}_t)}{\partial \mathbf{M}}$
  $t = t + 1$
end

Algorithm 1: An iterative gradient-descent algorithm to solve the optimization problems in Eqs. 5, 6, 7

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 [43], all shoes in our dataset 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 [43] for full implementation with minor modifications. In Fig. 7, 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 III-A, and the training procedure in Section IV-B, we produce a 13 dimensional binary attribute vector for each image and sketch in the dataset.

Parameters There are four parameters w_1, w_2, d, λ that need to be tuned in our model: w_1, w_2 weight different representations for late-fusion, d is dimension of the joint subspace, and λ trades off subspace and instance-alignment when learning our fine-grained subspace alignment. w_1, w_2, d and λ are fixed in all experiments, and empirically set to 0.3, 0.05, and 0.8, respectively. d is dataset dependent, and optimized using greedy search in the training set. We found 590, 400, 290, 190 and 890 to work well on 912-shoe dataset, 697-chair dataset, 419-Shoe [23], 297-Chair [23] and Sketchy [48], respectively.

Baselines (Attribute Detection): We compare our SS-DPM attribute detection to the conventional approach of using holistic features (**Whole-Image**), and using ground-truth part attributes as input (**Ground-Truth Part**). To verify that our part-aware method decorrelates the attributes, we evaluate against the state-of-the-art attribute decorrelation method introduced in [11], 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 [11], 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 [42] which works without strong part annotations at training (**Weakly-Supervised (WS)-DPM**).

Baselines (Cross-Modal Alignment): To evaluate our contribution to aligning low-level features across domains, we compare our proposed *fine-grained* subspace alignment (**FG-SA (Ours)**) with several prominent cross-modal alignment methods: **LCFS** [20], **BLM** [37], **CCA** [14], **GMMFA**, **GMLDA** [19], and **DA-SA** [13]. **DA-SA** is exactly the same as our domain-level subspace alignment method formulated in Eq. 1. **ILA** is our pair instance alignment method defined in Eq. 2. We adapt category-level methods (**LCFS, GMMFA, GMLDA**) to our problem by treating each sketch/image pair as an unique category label.

Baselines (Fine-grained SBIR): To evaluate our overall system combining high-level attribute and low-level feature alignment components, we evaluate the following variants. To evaluate representations, we consider **Part-HOG**, where part-level HOG is employed, **Part-Attribute**, where only automatically detected part-aware attributes are utilized, and **Part-Structure**, where geometric part structure alone is used to retrieve. To evaluate parts vs global encodings we consider: (i) **GLLF+Our**, where global HOG features are aligned by our proposed FG-SA method 1. (ii) **PLLF+CCA**, where part-level HOG features are projected into a common subspace using CCA. (iii) **PLLF+Our** where part-level HOG features are aligned by our proposed FG-SA method 1 and where only automatically detected part-aware attributes are utilized. Combining all three-views we have: **GLLF+Part-Attribute+Part-Structure+Our**, **PLLF+Part-Attribute+Part-Structure+Our**; which we compare to an earlier version of our work [47] denoted **PLLF+Part-Attribute+Part-Structure+CCA**.

B. Attribute Detection

In this section, we evaluate our attribute-detection performance on both domains. In Table I, we offer attribute detection accuracy (three-fold cross-validation for ten random splits) for 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. We can also see that: (i) All part-aware methods outperform whole-image, with the upper bound of ground-truth attributes offering the best performance. This justifies our contribution of part localization. (ii) Our method outperforms the state-of-the-art decorrelation method [11] on image and performs comparably on the more challenging sketch domain. Note that [11] required strong part annotations at testing, and our strongly-supervised DPM approach only used part annotation during training. (iii) SS-DPM outperforms the weakly-supervised alternative, again highlighting the importance of accurate part localization.

C. Cross-modal Matching with Low-Level Features

In this section, we evaluate our proposed *fine-grained* subspace alignment (FG-SA) with other prominent cross-modal or domain adaption methods on our proposed dataset, 419 Shoe dataset [23], 297 Chair dataset [23] and Sketchy [48]. Given each probe sketch, we retrieve K images, and define a successful retrieval if there is a correct match within those K images. From the CMC curves in Fig. 5(a),(b),(c),(d), we can see that (i) Our FG-SA which simultaneously learns cross-modality subspaces and aligns pair instances clearly outperforms the others across all four datasets, (ii) ILA already performs better than most category-level subspace alignment methods, suggesting the importance of pair instance alignment when it comes to *fine-grained* retrieval.

D. Fine-grained SBIR Performance Evaluation

Having evaluated our feature-alignment contribution, we next evaluate our entire SBIR framework including aligned low-level features and part-aware attributes (i.e., via Eq. (8)). The overall results are illustrated by CMC curve in Fig. 6, where we achieve an average of 60.89% @ $K = 10$, significantly outperforming the result in [47]. In particular, we make the observations: (i) The low-level feature alignment performance of our FG-SA (PLLF+Our 44.13% @ $K = 10$) is almost twice that of CCA (PLLF+CCA 25% @ $K = 10$), (ii) part-aware subspace projection improves the performance compared to global non-part aware projection (GLLF+Part-Attribute+Part-Structure+Our 54.19% @ $K = 10$ versus

Attribute	Whole-Image	WS-Decor [11]	WS-DPM	SS-Decor [11]	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-Image	WS-Decor [11]	WS-DPM	SS-Decor [11]	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 I
ATTRIBUTE DETECTION USING OUR PART-AWARE METHOD AND OTHER PREVIOUS STATE-OF-THE-ART METHODS ON BOTH IMAGE (LEFT) AND SKETCH (RIGHT) DOMAINS OUR METHOD GENERALLY PERFORMS BEST, WHERE SOME ATTRIBUTES ACTUALLY OUTPERFORM SS-DECOR. ONE EXCEPTION IS THAT ON SKETCH HEEL PART, WHERE SS-DECOR OUTPERFORMS OURS. NOTE HOWEVER THAT SS-DECOR REQUIRED STRONG PART ANNOTATION AT TESTING TIME, WHEREAS ONCE TRAINED OUR SS-DPMS WORK WITHOUT PART ANNOTATION AT TESTING.

PLLF+Part-Attribute+Part-Structure+Our 60.89% @ $K = 10$). In Fig. 7, 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 photos than the whole image approach by correctly matching the *fine-grained* details such as open vs. closed heel, or high-heel vs. platform.

E. Analysis on Different Drawing Styles

1) Sketching Styles: As shown in Fig. 3, different sketches completed by different annotators in our dataset have varying 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 our participants into six groups, where each group is made up of three individuals. Each group is 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 III, 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.

2) Sketchability of Attributes: Fine-grained Attributes are critical for *fine-grained* SBIR performance. However, not every annotator is a reliable sketcher and able to express the subtlety conveyed in images. Thus, in this section, we present a pilot study on how reliably users can convey *fine-grained* attributes via sketches. Specifically, we address this by digging into the attribute annotation on sketches and images in our dataset. Annotation in each domain was independently conducted by different workers. Assuming that workers can reliably annotate image attributes, the mismatch of sketch attribute annotations reflects the inability of users to sketch the attributes of the corresponding image. Table II shows the degree of annotation consensus across sketch-image pairs broken down by attribute, thus quantifying the sketchability of each attribute. These range from 76% at lowest (shoelace, pillar heel) to 97% at best (low-boot). In Fig. 8, we illustrate some challenging scenarios, where all the three users drawing the same shoe failed to express a given *fine-grained* attribute, according the annotators of the resulting sketches.

Round	Toe-open	Ornament or brand on body	Shoelace or ornament on vamp	Low heel	High heel	Pillar heel
94.41%	94.41%	83.77%	76.43%	86.73%	86.40%	76.43%
Cone heel	**Slender heel**	**Thick heel**	**Low boot**	**Middle boot**	**High boot**	**Average**
80.15%	92.10%	79.93%	97.36%	87.50%	90.13%	86.60%

TABLE II
PERCENTAGE OF ATTRIBUTES WHERE HUMAN ANNOTATORS AGREE ON SKETCH/IMAGE ATTRIBUTES. THIS SHOWS THAT SOME ATTRIBUTES ARE HARD TO INTERPRET IN SKETCHES, EVEN BY HUMANS; THUS ILLUSTRATING THE CHALLENGE OF THE *fine-grained* SBIR TASK.

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 III
***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.

F. Analysis on Different Drawing Styles

As shown in Fig. 3, 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 a *fine-grained* SBIR outcome. More specifically, at dataset generation, we divide our participants into six groups, where each group is made up of three individuals. Each group

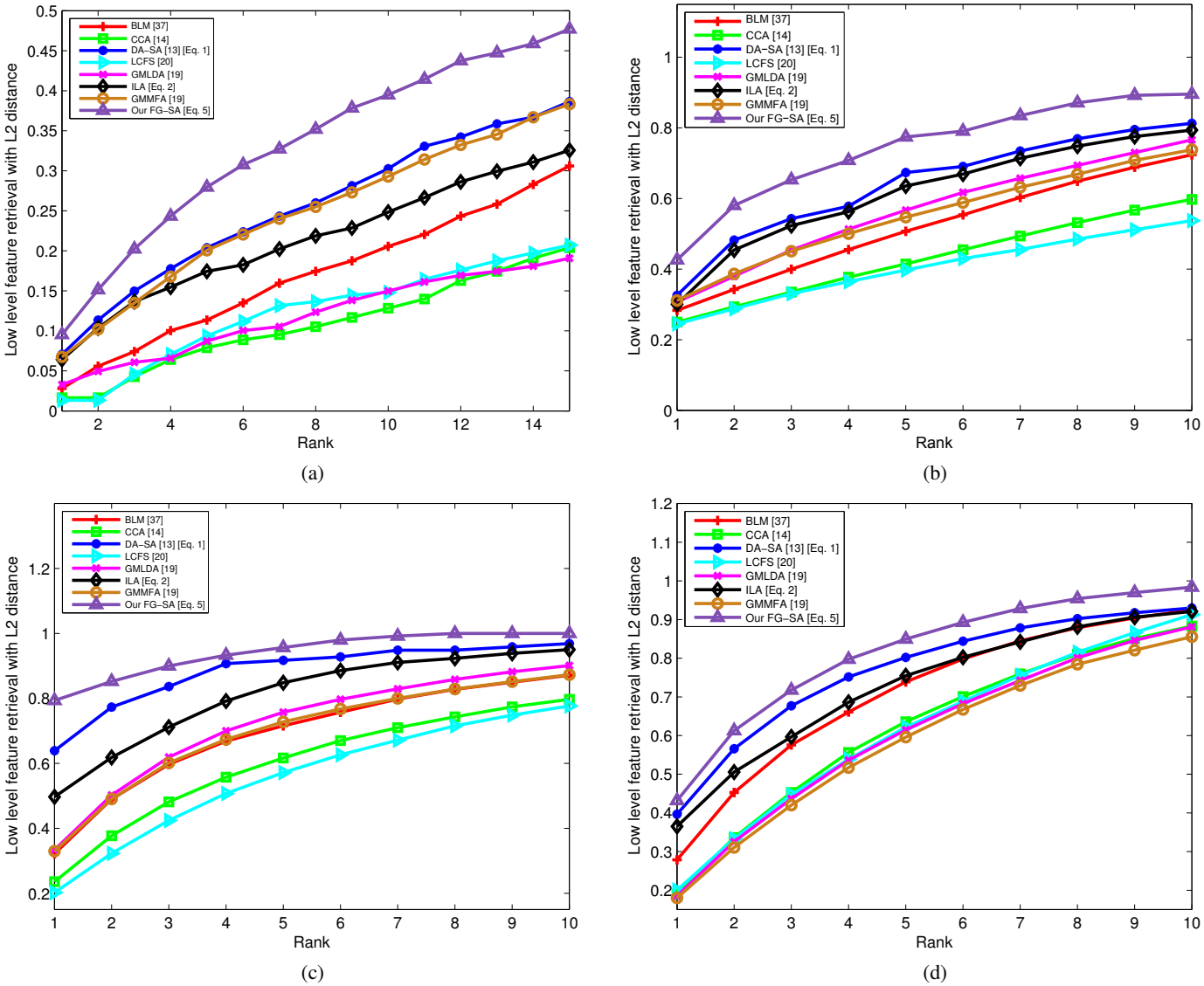


Fig. 5. CMC curves for *fine-grained* SBIR using low-level features only, compared with our proposed domain alignment method [Eq. 5] with alternatives on (a) 912 shoe sketch database, (b) 419 shoes database [23], (c) 297 chair database [23] and (d) Sketchy [48].

912 Shoe dataset [47]	**acc.@1**	**acc.@10**
Dense HOG + FG-SA-M1	9.84%	39.47%
Dense HOG + FG-SA-M2	8.88%	39.97%
Dense HOG + FG-SA-M3	8.22%	40.63%
ISN Deep + FG-SA-M1	8.55%	37.17%
ISN Deep + FG-SA-M2	**9.38%**	37.17%
ISN Deep + FG-SA-M3	8.39%	37.66%
Triplet Deep Learning [23]	8.06%	**43.42%**

TABLE IV
Fine-grained SBIR RESULTS. OUR *fine-grained* SUBSPACE ALIGNMENT METHODS VERSUS TRIPLET DEEP LEARNING: 912 SHOE SKETCH DATASET [47].

is 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 III, the performance of *fine-grained* SBIR can vary dramatically due to individuals' drawing styles. This result further highlights the challenging nature of the dataset and motivates future work to be carried out on user-specific modelling and/or developing style-robust models.

G. Further Evaluations on Alternative FG-SBIR Methods

Having previously evaluated against other cross-modal matching paradigms, in this set of experiments, we focus on (i) evaluating our three proposed variants for *fine-grained* subspace alignment, (ii) studying the cross-modal performance of our shallow approach compared with deep learning alternatives, and (iii) investigating our model performance against dataset size.

We compare with two deep learning approaches on their respective datasets where best results were reported: (i) a recently proposed deep triplet network for *fine-grained* SBIR

419 Shoe dataset [23]	acc.@1	acc.@10
Dense-HOG + FG-SA-M1	42.61%	89.57%
Dense-HOG + FG-SA-M2	42.61%	86.09%
Dense-HOG + FG-SA-M3	44.35%	87.83%
ISN Deep + FG-SA-M1	46.96%	89.57%
ISN Deep + FG-SA-M2	**51.30%**	90.43%
ISN Deep + FG-SA-M3	44.35%	**91.30%**
Triplet Deep Learning [23]	39.13%	87.83%

297 Chair dataset [23]	acc.@1	acc.@10
Dense-HOG + FG-SA-M1	**79.38%**	**100%**
Dense-HOG + FG-SA-M2	**79.38%**	**100%**
Dense-HOG + FG-SA-M3	71.13%	97.94%
ISN Deep + FG-SA-M1	76.29%	97.94%
ISN Deep + FG-SA-M2	76.29%	96.91%
ISN Deep + FG-SA-M3	71.13%	96.91%
Triplet Deep Learning [23]	69.07%	97.94%

TABLE V
Fine-grained SBIR RESULTS. OUR *fine-grained* SUBSPACE ALIGNMENT METHODS VERSUS TRIPLET DEEP LEARNING: 419 SHOE / 297 CHAIR DATASET [23].

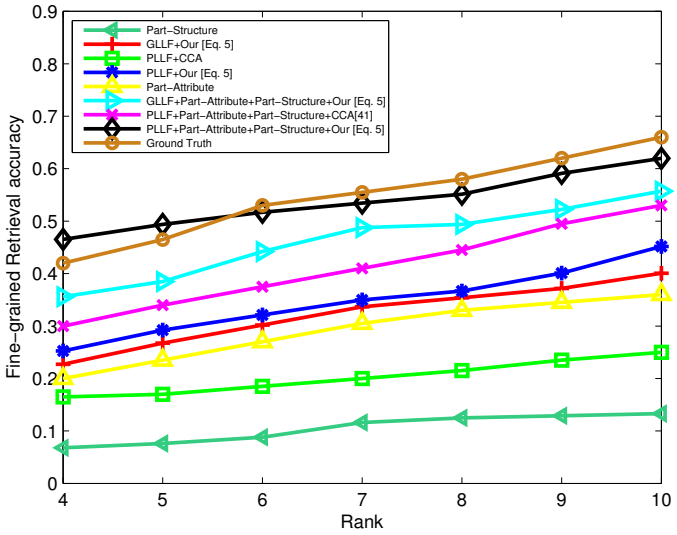


Fig. 6. CMC curves for the full *fine-grained* SBIR framework including part-aware attributes and domain alignment.

697 Chairs dataset	acc.@1	acc.@10
Dense HOG + FG-SA-M1	70.56%	97.97%
Dense HOG + FG-SA-M2	69.54%	97.97%
Dense HOG + FG-SA-M3	**72.08%**	**98.98%**
ISN Deep + FG-SA-M1	63.96%	91.88%
ISN Deep + FG-SA-M2	63.45%	93.40%
ISN Deep + FG-SA-M3	64.97%	93.91%
Triplet Deep Learning [23]	64.47%	90.36%

TABLE VI
Fine-grained SBIR RESULTS. OUR *fine-grained* SUBSPACE ALIGNMENT METHODS VERSUS TRIPLET DEEP LEARNING: NEW 697 CHAIR DATASET.

[23] that yields state-of-the-art results on 419-Shoe and 297-Chair [23]; and (ii) the heterogeneous triplet network based on GoogleNet proposed by [48] that delivers the best retrieval performance on the Sketchy dataset [48]. It is worth noting that, for direct comparison to [23] and [48], we do not use attributes and parts in this experimental section, nor utilize any of the exhaustively labeled triplets used to train the network in [23].

Sketchy [48]	acc.@1	acc.@10
GNT Deep + FG-SA-M1	42.85%	97.24
GNT Deep + FG-SA-M2	44.36%	97.18%
GNT Deep + FG-SA-M3	**45.27%**	**98.20%**
GN Triplet [48]	37.10%	95.53%

TABLE VII
Fine-grained SBIR RESULTS. OUR *fine-grained* SUBSPACE ALIGNMENT METHODS VERSUS GN TRIPLET: SKETCHY [48].

697 Chairs database To study the dependence of model performance on dataset size, we enlarge the previous chair database introduced by [23] to include an extra of 400 chair sketch-image pairs, making a new total of 697.

Low-Level Features In this experiment we use a holistic low-level feature representation obtained by densely extracting HOG on sketches and images on a 32×32 grid, resulting in a 2304 dimensional feature vector.

Baselines Our focus is on comparing our proposed 3 FG-SA methods with [23], which achieves the best *fine-grained* SBIR performance to date by employing a fine-tuned deep triplet network. Our FG-SA alignment operates on pre-extracted features and does not rely on triplet annotations, while deep learning method [23] operates on raw images and requires extensive triplet annotation.

Specifically, we evaluate the following cross-modal matching baselines: (i) **Triplet Deep learning**, which is exactly [23]. (ii) **Dense-HOG + Our**, where simple HOG features are aligned using our proposed FG-SA, and then compared using L2 distance. (iii) **ISN Deep + Our**, our FG-SA operates on pre-extracted ISN Deep features as also benchmarked in [23]. (iv) **GNT Deep + our**, our FG-SA operates on pre-extracted GoogleNet Triplet Deep features as proposed in [48].

The SBIR results on 912-dataset [47], 419-Shoe and 297-Chair [23], 697-dataset (introduced here) and Sketchy [48] are shown in Tables IV, V, VI, VII, respectively.

Comparison of the proposed three FG-SA methods We first compare our three method variants introduced in Section IV-C. As discussed earlier, one might expect that Method

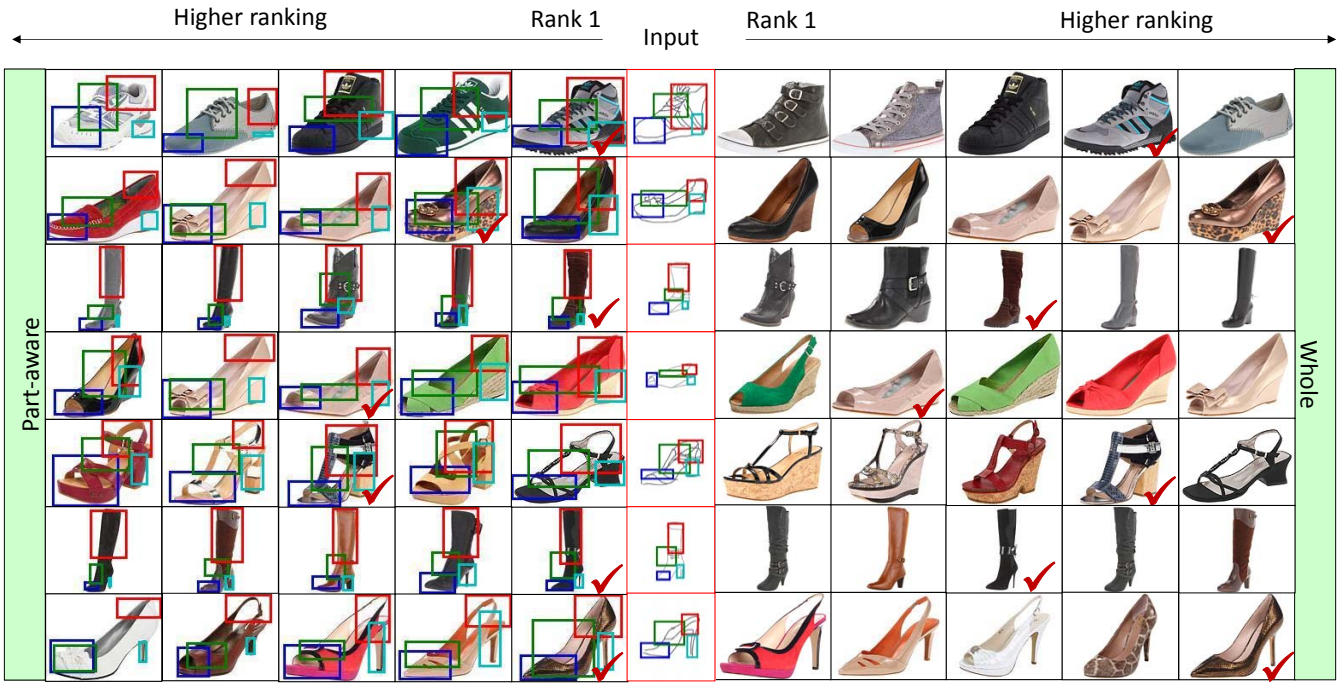


Fig. 7. **Fine-grained SBIR qualitative results with and without Part Decomposition** Examples of some top ranking retrieval results given a probe sketch. Our part-aware method delivers sensible results, discriminating the *fine-grained* variations on a instance-level. Red-tick indicates ground-truth matching image of the input sketch, which should be ranked as highly as possible. Part detection results of SS-DPM are shown using colour-coded bounding boxes.

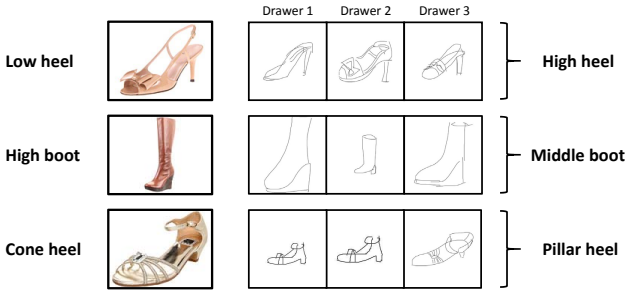


Fig. 8. Examples where humans fail to express a given image attribute when sketching. Given an image, all three sketchers failed to express the shown *fine-grained* attribute according to human annotations.

3 provides the strongest constraint (each positive pair closer than alternative negative pairs) for accurate cross-modal matching and should perform the best. However the results on the first two datasets in Tables V and IV do not support this, with Method 2 mostly performing best. This is because the tighter constraint is harder to meet, and thus the metric overfits to the training data while trying to meet it. In contrast, on the largest 697-dataset in Table VI, we do see that Method 3 performs best and Method 2 (positive pairs closer than negative pairs on average) is better than the Method 1 (positive pairs close). This suggests that stronger constraints are indeed effective, but only with sufficient training data.

Comparison against Deep Learning From the results we can observe that: There is no clear winner on the 912 shoe dataset (Table IV) – different outcomes at early and late ranks), our methods are comparable or better than Deep Triplet Ranking on the 419 shoe and 697 chair datasets (Tables V and VI), and beating GN Triplet on the Sketchy database (Tables VII). It is interesting to note that Dense-HOG features tend to perform better than ISN features on chairs, both on the 297-chair dataset of [23] and our extended 697-chair dataset. Upon close examination, we attribute this to (i) chair sketch-photo pairs exhibit much better alignment on the whole, and (ii) chair sketches are considerably better drawn than shoes (also reflected by the genially better top-1 accuracy and close to 100% performance on both chair datasets).

Annotation and Computational Cost Importantly, our method only requires pair correspondence, and not the non-scalable $O(N^3)$ triplet annotations need by [23]. Our FG-SA methods are also significantly more efficient to train. For example on the 419-shoe dataset [23]. It required 95 seconds, 628 seconds and $1,154$ seconds on average to train FG-SA Methods 1-3 respectively on CPU (i5-4590 @ 3.30GHz), compared to 9 hours on a Tesla-K80 GPU for [23].

VI. Conclusion

We investigated the practical problem of *fine-grained* sketch-based image retrieval. Our first contribution was to study the role of part-aware attributes. In doing so, we released 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. Our second contribution addressed the image/sketch divergence at the level of low-level features, by developing a *fine-grained* instance-level subspace alignment framework which, for the first time, aligns the modalities according to both domain-level and instance-level constraints. This novel approach compares favourably to more

expensive and complex deep learning approaches. Through comparative analysis against other existing subspace learning methods on *fine-grained* SBIR, we reinforced a conclusion drawn in a recent work [39], that is, for *fine-grained* tasks, pair-wise alignment is required. In the future, first we will investigate further the effect of style in *fine-grained* SBIR, and study automatic free-hand sketch synthesis from images [49], [50], [51]. We also plan to apply the proposed instance-level subspace alignment methods to other cross-modal matching tasks where pair correspondences are thought, e.g., person re-id [52], [53].

REFERENCES

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

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

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

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

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

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

[7] Y.-L. Lin, C.-Y. Huang, H.-J. Wang, and W.-C. Hsu, "3d sub-query expansion for improving sketch-based multi-view image retrieval," in *ICCV*, 2013.

[8] 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.

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

[10] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *CVPR*, 2014.

[11] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *CVPR*, 2014.

[12] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*, 2012.

[13] B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, K. U. Leuven, L. Hubert, C. Umr, and B. Lauras, "Unsupervised Visual Domain Adaptation Using Subspace Alignment," *ICCV*, 2013.

[14] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," *ACM Multimedia*, 2010.

[15] T. K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *TPAMI*, 2007.

[16] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *CVPR*, 2011.

[17] C. Dhanjal, S. R. Gunn, and J. Shawe-Taylor, "Efficient sparse kernel feature extraction based on partial least squares," *TPAMI*, 2008.

[18] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intraindividual correlations for face recognition across pose differences," 2009.

[19] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: a discriminative latent space," *CVPR*, 2012.

[20] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning Coupled Feature Spaces for Cross-Modal Matching," *ICCV*, 2013.

[21] R. Zhou, L. Chen, and L. Zhang, "Sketch-based image retrieval on a large scale database," in *ICMR*, 2012.

[22] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014.

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

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

[25] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.

[26] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus, "Learning invariance through imitation," in *CVPR*, 2011.

[27] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from flickr groups using stochastic intersection kernel machines," in *ICCV*, 2009.

[28] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011.

[29] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014.

[30] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *ECCV*, 2010.

[31] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[32] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009.

[33] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *CVPR*, 2014.

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," tech. rep., 2011.

[35] 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*, 2014.

[36] 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*, 2015.

[37] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, 2000.

[38] M. Long, J. Wang, G. Ding, J. Sun, and P. Yu, "Transfer joint matching for unsupervised domain adaptation," in *CVPR*, 2014.

[39] P. Xu, Q. Yin, Y. Huang, Y.-Z. Song, Z. Ma, L. Wang, T. Xiang, W. B. Kleijn, and J. Guo, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *http://arxiv.org/abs/1705.09888*, 2017.

[40] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *ECCV*, 2010.

[41] A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: Image search with relative attribute feedback," in *CVPR*, 2012.

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

[43] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *ECCV*, 2012.

[44] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *NIPS*, 2010.

[45] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012.

[46] H. Bischof, P. M. Roth, M. Hirzer, P. Wohlhart, and M. Kostinger, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[47] K. Li, K. Pang, Y.-Z. Song, T. Hospedales, H. Zhang, and Y. Hu, "Fine-grained sketch-based image retrieval: The role of part-aware attributes," in *WACV*, 2016.

[48] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *Acm Transactions on Graphics*, 2016.

[49] N. Wang, X. Gao, L. Sun, and J. Li, "Bayesian face sketch synthesis," *TIP*, 2017.

[50] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketchphoto synthesis and retrieval using sparse representation," *TCSVT*, 2012.

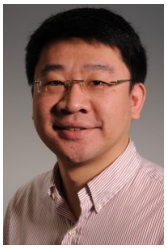
[51] Y. Li, Y.-Z. Song, T. M. Hospedales, and S. Gong, "Free-hand sketch synthesis with deformable stroke models," *International Journal of Computer Vision*, vol. 122, no. 1, pp. 169–190, 2017.

[52] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[53] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person reidentification by one-shot group-based verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 591–606, 2016.

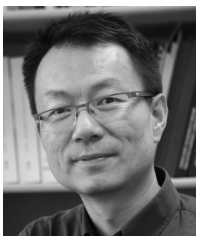
Ke Li is currently a Ph.D candidate at School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. His research interest is computer vision, particularly focus on fine-grained sketch-based on image retrieval.

Kaiyue Pang is currently a Ph.D candidate at SketchX Research Lab, Queen Mary University of London. His research interest is computer vision, particularly focus on generaitve and discriminative modelling of human sketches and how such can be transferred into novel commercial applications.



Yi-Zhe Song received the Ph.D degree in Computer Vision and Graphics from the University of Bath in 2008. He is currently a Senior Lecturer (Associate Professor), and the founding Director of SketchX Research Lab in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests lay solely on understanding human sketches and enabling novel applications of sketches. He has published more than 50 papers in major international journals and conferences. He is a member of the IEEE.

Timothy Hospedales received the Ph.D degree in neuroinformatics from the University of Edinburgh in 2008. He is currently a Reader (associate professor) at The University of Edinburgh. His research interests include probabilistic modeling and machine learning,particularly life-long, transfer, and active learning with applications to computer vision and beyond.



Tao Xiang received the Ph.D degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 140 papers in international journals and conferences.

Honggang Zhang received the B.S degree from the department of Electrical Engineering, Shandong University in 1996, the Master and Ph.D degrees from the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT) in 1999 and 2003 respectively. He worked as a Visiting Scholar in School of Computer Science, Carnegie Mellon University (CMU) from 2007 to 2008. He is currently an Associate Professor and Director of web search center at BUPT. His research interests include image retrieval, computer vision and pattern recognition. He published more than 30 papers on TPAMI, SCIENCE, Machine Vision and Applications, AAAI, ICPR, ICIP. He is a senior member of IEEE.