Instance-level Coupled Subspace Learning for Fine-grained Sketch-based Image Retrieval

Peng Xu[1], Qiyue Yin[2], Yonggang Qi[1], Yi-Zhe Song[3], Zhanyu Ma[1]⋆,
Liang Wang[2], Jun Guo[1]

[1]Pattern Recognition and Intelligent Systems Lab., Beijing University of Posts and Telecommunications, Beijing, China.
[2]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[3]SketchX Lab, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom.
{peng.xu, qiyg, mazhanyu, guojun}@bupt.edu.cn {qyyin, wangliang}@nlpr.ia.ac.cn yizhe.song@qmul.ac.uk

Abstract. Fine-grained sketch-based image retrieval (FG-SBIR) is a newly emerged topic in computer vision. The problem is challenging because in addition to bridging the sketch-photo domain gap, it also asks for instance-level discrimination within object categories. Most prior approaches focused on feature engineering and fine-grained ranking, yet neglected an important and central problem: how to establish a fine-grained cross-domain feature space to conduct retrieval. In this paper, for the first time we formulate a cross-domain framework specifically designed for the task of FG-SBIR that simultaneously conducts instance-level retrieval and attribute prediction. Different to conventional photo-text cross-domain frameworks that performs transfer on category-level data, our joint multi-view space uniquely learns from the instance-level pair-wise annotations of sketch and photo. More specifically, we propose a joint view selection and attribute subspace learning algorithm to learn domain projection matrices for photo and sketch, respectively. It follows that visual attributes can be extracted from such matrices through projection to build a coupled semantic space to conduct retrieval. Experimental results on two recently released fine-grained photo-sketch datasets show that the proposed method is able to perform at a level close to those of deep models, while removing the need for extensive manual annotations.

Keywords: Fine-Grained SBIR, Attribute Supervision, Attribute Prediction, Multi-view domain adaptation.

1 Introduction

Sketch-based image retrieval (SBIR) is traditionally casted into a classification problem, and most prior art evaluates retrieval performance at category-level. [1–10], i.e. given a query sketch, the goal is to discover photos with the same class

⋆ Corresponding author.

label. However, it was recently argued [11, 12] that SBIR is more reasonable to be conducted at a fine-grained level, where instead of conducting retrieval across object categories, it focuses on finding similar photos to the query sketch within specific categories. By specifically exploring the unique fine-grained visual characteristics captured in human sketches, fine-grained SBIR is likely to transform the traditional landscape of image retrieval by introducing a new form of user interaction that underpins the ubiquitous commercial adoption of SBIR technology.

Shared with conventional category-level SBIR, the core problem of fine-grained SBIR lies with that of cross-domain, that is sketches and photos are from inherently heterogeneous domains. This domain difference can be summarized into two main gaps: (i) the visual modality gap, i.e., sketches are coarse line drawings with plain white background and photos are made of dense color pixels on textured background, and (ii) the semantic gap, i.e., free-hand sketches are highly abstract and iconic, whereas photos are pixel-perfect depictions of the visual world. The problem is further made difficult for fine-grained SBIR since *fine-grained* correspondence between sketch and photo is difficult to establish especially given the abstract and iconic nature of free-hand sketches. It is therefore important for any fine-grained SBIR framework to not only seek a fine-grained metric, but also learn a joint semantic space to effectively model the domain gap.

Prior work on fine-grained SBIR either focused on feature engineering [11] or learning a fine-grained feature space [12]. There has been a largely neglected problem of addressing the cross-domain gap per sa. Majority of work ease the domain gap by first converting images to edgemaps, and conduct further comparisons by treating the extracted edgemaps as somewhat “good” sketches. For example, Yu et al. employed Sketch-a-Net [13] that is specifically designed to parse sketches for both photo and sketch branches in their triplet ranking network. However, sketches and photos are fundamentally different: photos closely follow natural image statistics and are taken by cameras, yet sketches are drawn from visual memory and produced by hand. In this work, for the first time, we explicitly model the cross-domain gap between photo and sketch by jointly learning a coupled semantic embedding using fine-grained visual attributes.

Parallel to traversing the photo-sketch domain gap, the modality gap between text and photo has been widely studied in recent years [14–20]. In essence, the goal of cross-modal techniques is to shorten the semantic gap between text and photo through projecting the inherently different domains into a common subspace and consequently perform matching. Although many were shown to able to effectively traverse the cross-domain gap, they only conduct transfer at category-level or domain-level, rendering them unsuitable for fine-grained retrieval where instance-level differences are sought after instead. Our cross-domain model on the other hand learns from instance-level sketch-photo pairs, resulting in a subspace that is not only domain-independent, but also fine-grained.

In this paper, we present a novel subspace learning method for FG-SBIR based on attribute supervision and view selection. Our framework performs join-

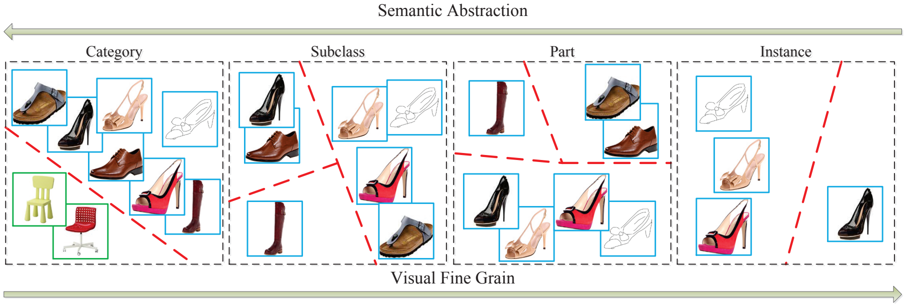


Fig. 1. Retrievals based on different level of grains. The top arrow from right to left denotes the enhanced semantic abstraction. The bottom arrow from left to right indicates increasing fine-grained level.

t attribute regressions for sketch and photo modalities, which is able to select relevant and discriminative feature views from coupled sketch-photo spaces simultaneously. The goal is to project sketch and photo features into coupled attribute spaces. Meanwhile, such space is also capable of predicting attributes by multiplying the learned projection matrices. Specifically, our objective function consists of three parts: (i) coupled supervised linear regression, (ii) coupled group norms of all projection matrices, and (iii) a Frobenius norm regularization. The coupled supervised linear regressions take advantage of the rich attribute information to learn local feature-wise relationships at an abstract level. The group norms of the projection matrices play the role of simultaneous and joint view selection among multi-view features. The Frobenius norm regularization can bridge the gap between sketch-photo attribute spaces. Accordingly, an efficient algorithm is derived to solve the proposed optimization problem. Experimental results on two fine-grained image-sketch datasets demonstrate that the proposed method outperforms the state-of-the-art shallow approaches and its performance is even close to the deep models.

The main contributions of our work are as follows:

1. We propose for the first time an unified cross-domain framework of FG-SBIR.
2. We study how fine-grained visual attributes can be useful to construct a fine-grained and domain-independent joint feature space
3. We introduce an efficient algorithm to solve the challenging non-smooth optimization problem.
4. The proposed method outperforms state-of-the-art shallow models and offers comparable performance against deep alternatives on two recently released fine-grained photo-sketch datasets.

2 Related Work

SBIR vs. Fine-grained SBIR Traditional sketch-based retrieval tasks usually focus on global visual similarities and high-level semantics. As a result, retrieval is often performed coarsely at category-level. In contrast, fine-grained retrieval paradigms concentrate on subtle visual and semantic descriptions of objects. As shown in Fig. 1, most SBIR work can be broadly summarized into four categories according to the level of detail they operate on: (i) *Category-level retrieval* aims to examine objects on category-level [3, 4, 10], e.g., shoes against chairs; (ii) *Subclass-level retrieval* differentiate objects on within-class category level, e.g., shoes are classified into three subcategories according to their general usage; (iii) *Part-level retrieval* finds objects according to the subtle part properties [21], e.g. four high-heel shoes are marked out according to the properties of heel and boot; (iv) For *fine-grained instance-level retrieval* [11, 12], the sketch shoe and two high-heel sandals become the nearest neighbors on the basis of similarities on the heel, body, and toe. Our proposed fine-grained SBIR model is able to generalize to all four variations, and we offer experimental comparisons for each later in Section 4.

Towards Fine-grained SBIR Li et al. [11] first proposed fine-grained SBIR (FG-SBIR) but limited their study to pose variations only and the cross-domain gap is only traversed holistically by matching coarse graph structures. Yu et al. [12] further extended the definition of fine-grained and proposed a new dataset of sketch-photo pairs with detailed triplet annotation. They developed a deep triplet-ranking network to learn a fine-grained feature metric, however avoided addressing the cross-domain gap by converting photos to edgemaps prior to training and testing. The very recent work of Li et al. [21] remains the single work that specifically tackled the cross-domain nature of the problem, where they used three-view Canonical Correlation Analysis (CCA) to fuse fine-grained visual attributes and low-level features. However, they did not learn a joint feature space since CCA is only conducted independently on each domain. Moreover, it required separately trained set of attribute detectors at testing time, making it less generalizable to other datasets. In this paper, we follow Li et al. [21] in using fine-grained attributes to traverse different domains, but explicitly learn a joint fine-grained space to conduct retrieval. Once learned, this attribute-driven space is also able to perform implicit attribute detection without additional training.

Cross-modal Retrieval Broadly speaking, cross-modal retrieval involves two main tasks: measure of relevance and coupled feature selection [14]. The challenge of cross-modal matching is therefore finding a semantic feature space that can withstand modal variation at an abstract level. Most cross-modal methods can be classified into three main categories: probabilistic models [15, 16], metric learning approaches [17, 18] and subspace learning methods [19, 20]. Probabilistic approaches aim to model the joint distribution of multi-modal data in order to learn their correlation [15]. Metric learning methods set out to compute appropriate distance metrics between different modalities [17]. Subspace learning approaches map multi-modal data into a common subspace to conduct matching [14]. Among these categories of cross-modal techniques, subspace learning methods [22–24, 14] have gained state-of-the-art results in recent years. All aforementioned cross-domain models can not work with instance-level annotations (e.g., sketch-photo pairs), largely limiting their applicability for fine-grained retrieval. Our proposed model is however specifically designed to mine a joint subspace where cross-domain comparisons can be performed at a fine-grained level.

3 Fine-Grained SBIR via Attribute Supervision and View Selection

In this section, we introduce our framework for FG-SBIR based on attribute supervision and view selection. An effective algorithm is also presented to solve the proposed objective function.

3.1 Notations

Matrices and column vectors will be consistently denoted as bold uppercase letters and bold lowercase letters, respectively. Given a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, we express its i-th row as $\mathbf{M}^i$ and j-th column as $\mathbf{M}_j$.

The Frobenius norm of the matrix $\mathbf{M}$ is defined as

$$\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{m} \|\mathbf{M}^i\|_2^2} \ . \tag{1}$$

The Group ℓ_1-norm (G_1-norm) of the matrix $\mathbf{M}$ is defined as

$$\|\mathbf{M}\|_{G_1} = \sum_{i=1}^{n}\sum_{j=1}^{k} \|\mathbf{m}_i^j\|_2 \ , \tag{2}$$

where $\mathbf{m}_i^j$ is the j-th segment vector in the i-th column of $\mathbf{M}$.

3.2 Problem Formulation

Suppose there are n pairs of photo and sketch, which are denoted as $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n] \in \Re^{d^p\times n}$ and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n] \in \Re^{d^s\times n}$, respectively. As illustrated in Fig. 2, $\mathbf{p}_i \in \Re^{d^p}$ is formed by stacking features from all the k^p views, and the feature for each view j is a d_j^p dimensional vector, i.e. $d^p = \sum_{j=1}^{k^p} d_j^p$, similarly so for each element $\mathbf{s}_i$ in $\mathbf{S}$. The features used for different views can be low-level features (e.g., HOG), or those extracted from deep networks, (e.g., [12]). Each photo-sketch pair $\{\mathbf{p}_i, \mathbf{s}_i\}$ represents the same object. Let $\mathbf{A}_p = [\mathbf{a}_1^p, \mathbf{a}_2^p, ..., \mathbf{a}_n^p]^T \in \Re^{n\times u}$ denotes the attribute label matrix of the photo samples and u is the number of photo attribute. Similarly, $\mathbf{A}_s = [\mathbf{a}_1^s, \mathbf{a}_2^s, ..., \mathbf{a}_n^s]^T \in \Re^{n\times v}$ denotes the attribute label matrix of the sketch samples and v is the number of sketch attribute.

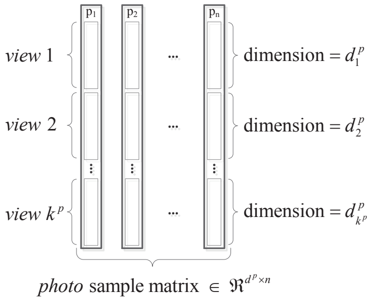


Fig. 2. Illustration of the photo sample matrix, **P**.

As previously discussed, SBIR and FG-SBIR generally belong to the task of cross-modal retrieval. Recently, many cross-modal approaches [22–25, 14, 26, 27] have achieved satisfying results on matching photo and text. Yet, all of them evaluated retrieval results on category-level by calculating the mean average precision (MAP) [28]. More specifically, given multi-modal sample matrices $\mathbf{X}_a$, $\mathbf{X}_b$, and class label matrix $\mathbf{Y}$, we can summarize a framework for supervised cross-modal subspace learning:

$$\min_{\mathbf{W}_a, \mathbf{W}_b} \|\mathbf{X}_a^T \mathbf{W}_a - \mathbf{Y}\|_F^2 + \|\mathbf{X}_b^T \mathbf{W}_b - \mathbf{Y}\|_F^2 + \Omega \ , \tag{3}$$

where $\mathbf{W}_a$ and $\mathbf{W}_b$ are the projection matrices and Ω is some form of constraint.

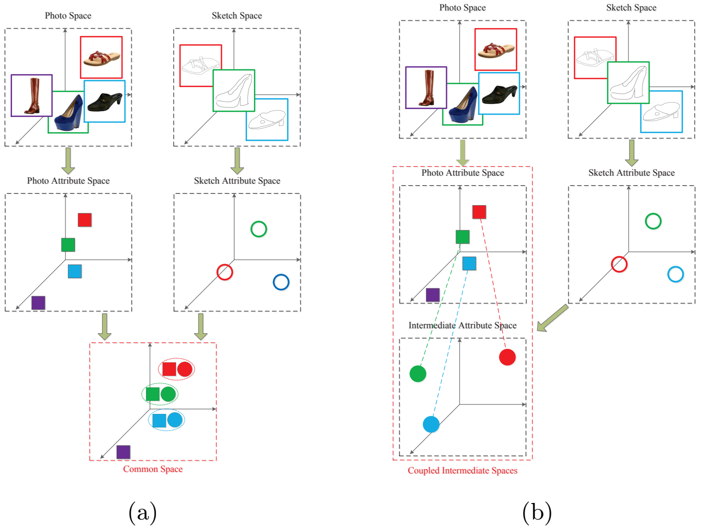
In this paper, we would like to conduct FG-SBIR in the visual attribute spaces. It follows that Eq. (3) naturally inspires us to project sketch and photo into a common attribute subspace as shown in Fig. 3(a). However, it would otherwise be difficult to define or annotate a desired common space and give it a clear semantic interpretation like the low dimensional class label matrix Y used in usual cross-modal frameworks. Motivated by several unsupervised cross-modal subspace learning methods [22–25], we propose to map sketch and photo data into two intermediate and isomorphic spaces U^S and U^P that have a natural correspondence. This means that U^S and U^P are approximation versions for each other in the ideal case. It follows that we can establish invertible mappings as follows:

$$\Re^{d^p} \rightleftarrows U^P \rightleftarrows U^S \rightleftarrows \Re^{d^s} \ . \tag{4}$$

The photo attribute space $\Re^u$ itself can potentially be directly used as its intermediate space U^P as shown in Fig. 3(b). For constructing the intermediate space of sketch U^S, the following can be adopted to approach U^P:

$$U^P \longleftarrow \mathbf{A}_s \mathbf{T}_s \ , \quad U^P \longleftarrow \mathbf{A}_p \mathbf{T}_p \ . \tag{5}$$

where $\mathbf{T}_s$ and $\mathbf{T}_p$ are the transformation matrices for sketch sample attribute matrix $\mathbf{A}_s$ and photo sample attribute matrix $\mathbf{A}_p$, respectively. Mathematically, we have $\min\limits_{\mathbf{T}_s} \|\mathbf{A}_p - \mathbf{A}_s \mathbf{T}_s\|_F^2$, and $\min\limits_{\mathbf{T}_p} \|\mathbf{A}_p - \mathbf{A}_p \mathbf{T}_p\|_F^2$.



(a) (b)

Fig. 3. Schematic comparison of conventional common-space learning (a), and the proposed coupled space learning (b)

An important point to note here is that as a result of the abstract nature sketches, they are often harder to interpret, resulting in a higher degree of noise in human attribute annotation when compared with photos. Hence the sketch sample attribute matrix $\mathbf{A}_s$ often loses information and is stuck in sparsity and low rank. For these reasons, in practice, we opt to the following to approach U^P: $\min_{\mathbf{T}_p} \|\mathbf{A}_p - \mathbf{A}_p\mathbf{T}_p\|_F^2$, whose optimization process starts from $\mathbf{A}_p$.

Our goal is to learn two projection matrices $\mathbf{W}_p$ and $\mathbf{W}_s$ jointly to map the associated data pairs into coupled intermediate spaces denoted by the corresponding attribute labels, subject to that the distance should be small if they belong to the same object. Therefore, the proposed objective function is formulated as follows:

$$\begin{aligned} J = \min_{\mathbf{W}_p,\mathbf{W}_s,\mathbf{T}} & \|\mathbf{P}^T\mathbf{W}_p - \mathbf{A}_p\|_F^2 + \|\mathbf{S}^T\mathbf{W}_s - \mathbf{A}_p\mathbf{T}\|_F^2 \\ & + \lambda_1(\|\mathbf{W}_p\|_{G_1} + \|\mathbf{W}_s\|_{G_1}) + \lambda_2\|\mathbf{A}_p - \mathbf{A}_p\mathbf{T}\|_F^2 \ , \end{aligned} \tag{6}$$

where $\mathbf{W}_p \in \Re^{d^p \times u}$ and $\mathbf{W}_s \in \Re^{d^s \times u}$ are the projection matrices for coupled photo and sketch spaces, respectively. $\mathbf{W}_p$ is a matrix which consist of weights for features from each individual view over u different attributes. And $\mathbf{W}_p$ can be re-written as:

$$\mathbf{W}_p = \begin{bmatrix} (\mathbf{w}_1^p)^1 & (\mathbf{w}_2^p)^1 & \cdots & (\mathbf{w}_u^p)^1 \\ (\mathbf{w}_1^p)^2 & (\mathbf{w}_2^p)^2 & \cdots & (\mathbf{w}_u^p)^2 \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{w}_1^p)^{k^p} & (\mathbf{w}_2^p)^{k^p} & \cdots & (\mathbf{w}_u^p)^{k^p} \end{bmatrix}, \tag{7}$$

where $(\mathbf{w}_x^p)^y \in \Re^{d_y^p}$ is a weighting vector contains the weights for all features in the y-th view of p (photo) sample with respect to the x-th attribute. $\mathbf{T} \in \Re^{u \times u}$ is a conversion matrix.

Similarly:

$$\mathbf{W}_s = \begin{bmatrix} (\mathbf{w}_1^s)^1 & (\mathbf{w}_2^s)^1 & \cdots & (\mathbf{w}_v^s)^1 \\ (\mathbf{w}_1^s)^2 & (\mathbf{w}_2^s)^2 & \cdots & (\mathbf{w}_v^s)^2 \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{w}_1^s)^{k^s} & (\mathbf{w}_2^s)^{k^s} & \cdots & (\mathbf{w}_v^s)^{k^s} \end{bmatrix} . \tag{8}$$

We want to present sketch data in an approximate space of the photo attribute space. By minimizing the projected residuals with respect to attribute information, we can preliminarily shorten the gap between the coupled intermediate spaces. And we can minimize the term $\lambda_2 \|\mathbf{A}_p - \mathbf{A}_p\mathbf{T}\|_F^2$ to learn the relationship $\mathbf{T}$ between the coupled attribute intermediate spaces. $\mathbf{T}$ contains the attribute mappings across U^S and U^P.

$\mathbf{W}_p$ and $\mathbf{W}_s$ are able to learn the weight vector for each single view feature, such that the feature-wise importance corresponding to a certain attribute in the intermediate spaces can be captured. However, the multi-view features interactions are extremely complicated, i.e., inhibition, promotion or competition depending on differnet cases. To solve this problem, motivated by [29], a Group ℓ_1-norm (G_1-norm) is utilized, i.e., the second part of Eq. (6).

According to the effectiveness of paired Group ℓ_1-norms upon $\mathbf{W}_p$ and $\mathbf{W}_s$, inside each column of these two projection matrices, the weight vectors for multi-view features are organized under the ℓ_1-norm framework. The view-wise relationships of ℓ_1-norm enforces the structured sparsity among different views. If certain view of features does not own enough contribution or discrimination for certain attribute, the corresponding weight vector of this view will be assigned with zeros, and vice versa. Within each column inside photo or sketch modality, the local interrelations among views are captured by Group ℓ_1-norm regularizer.

More importantly, our objective function optimizes the Group ℓ_1-norm regularizers of $\mathbf{W}_p$ and $\mathbf{W}_s$ simultaneously. Therefore, multi-modal data is fully integrated and equally taken into account to complete more reasonable view selection without unnecessary information loss. All the weight vectors for all the views are organized under the ℓ_1-norm framework. Hence the global relationships among all the views are also captured by the coupled Group ℓ_1-norm regularizers:

$$\begin{aligned} \|\mathbf{W}_p\|_{G_1} + \|\mathbf{W}_s\|_{G_1} &= \sum_{i=1}^{u}\sum_{j=1}^{k^p} \|(\mathbf{w}_p)_i^j\|_2 + \sum_{i=1}^{u}\sum_{j=1}^{k^s} \|(\mathbf{w}_s)_i^j\|_2 \\ &= \sum_{i=1}^{u}(\sum_{j=1}^{k^p} \|(\mathbf{w}_p)_i^j\|_2 + \sum_{j=1}^{k^s} \|(\mathbf{w}_s)_i^j\|_2) \ . \end{aligned} \tag{9}$$

In summary, the residual terms based on the attribute labels use the semantic information to preliminarily shorten the gaps between photo-sketch pairs across the coupled intermediate spaces. Next the Group ℓ_1-norm terms captured the local interrelations of multi-view features inside photo or sketch and the global relationships of data pairs crossing photo and sketch modalities. Finally the Frobenius norm term enforces the accuracy of attribute space transition.

3.3 Solving for Non-smooth Optimization

The designed objective function contains the non-smooth regularization terms of Group ℓ_1-norm, which is difficult to solve by general methods. The unknown quantities of our objective function are $\mathbf{W}_p$, $\mathbf{W}_s$, and $\mathbf{T}$. Fortunately, our objective function has no constraint conditions. We can use the variable separation approach to derive an alternative iterative algorithm to solve it.

Take the derivative of the objective J with respect to $(\mathbf{W}_p)_i$ $(1 \le i \le u)$, we have [1]

$$\frac{\partial J}{\partial (\mathbf{W}_p)_i} = 2\mathbf{P}\mathbf{P}^T(\mathbf{W}_p)_i - 2\mathbf{P}(\mathbf{A}_p)_i + \lambda_1 \mathbf{D}_p^i (\mathbf{W}_p)_i \ , \tag{10}$$

where $\mathbf{D}_p^i$ is a block diagonal matrix with the j-th diagonal block as $\frac{1}{2\|(\mathbf{W}_p)_i^j\|_2}\mathbf{I}_j$, $\mathbf{I}_j$ is an identity matrix with the same size as d_j^p, $(\mathbf{W}_p)_i^j$ is the j-th segment of $(\mathbf{W}_p)_i$ and includes the weighting vector for the features in the j-th view of photo sample matrix. Set $\frac{\partial J}{\partial (\mathbf{W}_p)_i} = 0$, we can get

$$(\mathbf{W}_p)_i = (2\mathbf{P}\mathbf{P}^T + \lambda_1 \mathbf{D}_p^i)^{-1}(2\mathbf{P}(\mathbf{A}_p)_i) \ . \tag{11}$$

Similarly, we can obtain $(\mathbf{W}_s)_i$ as

$$(\mathbf{W}_s)_i = (2\mathbf{S}\mathbf{S}^T + \lambda_1 \mathbf{D}_s^i)^{-1}(2\mathbf{S}\mathbf{A}_P(\mathbf{T})_i) \ . \tag{12}$$

Take the derivative of the objective J with respect to $(\mathbf{T})_i$ $(1 \le i \le u)$, and set $\frac{\partial J}{\partial (\mathbf{T})_i} = 0$, we can get

$$(\mathbf{T})_i = (\mathbf{A}_p^T\mathbf{A}_p + \lambda_2 \mathbf{A}_p^T\mathbf{A}_p)^{-1}(\mathbf{A}_p^T\mathbf{S}^T(\mathbf{W}_s)_i + \lambda_2 \mathbf{A}_p^T(\mathbf{A}_p)_i) \ . \tag{13}$$

Note that $\mathbf{D}_p^i$ $(1 \le i \le u)$ and $\mathbf{D}_s^i$ $(1 \le i \le u)$ are dependent on $\mathbf{W}_p$ and $\mathbf{W}_s$, respectively. We can optimize them alternatively and iteratively until convergence. During each optimization step of $\mathbf{W}_p$, $\mathbf{W}_s$, and $\mathbf{T}$, both of them are obtained column by column.

The whole algorithm is summarized in Algorithm 1.

[1] When $\|(\mathbf{W}_p)_i^j\|_2 = 0$, (6) is not differentiable. Following [30], a small perturbation can be introduced to smooth the j-th diagonal block of $\mathbf{D}_p^i$ as $\frac{1}{2\sqrt{\|(\mathbf{W}_p)_i^j\|_2^2 + \zeta}}\mathbf{I}_j$. Similarly, when $\|(\mathbf{W}_s)_i^j\|_2 = 0$, the j-th diagonal block of $\mathbf{D}_s^i$ can be regularized as $\frac{1}{2\sqrt{\|(\mathbf{W}_s)_i^j\|_2^2 + \zeta}}\mathbf{I}_j$. We set $\zeta = 1.0000e - 8$ in our following experiments.

Algorithm 1 An efficient iterative algorithm to solve the optimization problem in Eq. (6).

```
Input: $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n] \in \Re^{d^p \times n}$, $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n] \in \Re^{d^s \times n}$,
       $\mathbf{A}_p = [\mathbf{a}_1^p, \mathbf{a}_2^p, ..., \mathbf{a}_n^p]^T \in \Re^{n \times u}$.
  1.Set $t = 0$.
  Initialize $(W_p)_t$, $(W_s)_t$ by solving $\min_{W_p} \|P^T W_p - A_p\|_F^2$ and $\min_{W_s} \|S^T W_s - A_p T\|_F^2$ respectively.
  Initialize $(T)_t$.
  while not converge do
      2.Calculate the block diagonal matrices $(D_p^i)_{t+1}$ $(1 \leqslant i \leqslant u)$ and $(D_s^i)_{t+1}$ $(1 \leqslant i \leqslant u)$,
      where the j-th diagonal block of $(D_p^i)_{t+1}$ is $\frac{1}{2\|((W_p)_i^j)_t\|_2} I_j$
      and the j-th diagonal block of $(D_s^i)_{t+1}$ is $\frac{1}{2\|((W_s)_i^j)_t\|_2} I_j$.
      3.For each $(W_p)_i$ $(1 \leqslant i \leqslant u)$,
      $((W_p)_i)_{t+1} \leftarrow (2PP^T + \lambda_1 (D_p^i)_{t+1})^{-1} (2P(A_p)_i)$.
      4.For each $(W_s)_i$ $(1 \leqslant i \leqslant u)$,
      $((W_s)_i)_{t+1} \leftarrow (2SS^T + \lambda_1 (D_s^i)_{t+1})^{-1} (2SA_p(T_i)_t)$.
      5.For each $(T)_i$ $(1 \leqslant i \leqslant u)$,
      $(T_i)_{t+1} \leftarrow (A_p^T A_p + \lambda_2 A_p^T A_p)^{-1} (A_p^T S^T ((W_s)_i)_{t+1} + \lambda_2 A_p^T (A_p)_i)$.
      6.$t \leftarrow t + 1$.
  end while
Output: $W_p \in \Re^{d^p \times u}$, $W_s \in \Re^{d^s \times u}$, and $T \in \Re^{u \times u}$.
```

4 Experimental Results and Discussions

In this section, we describe how to apply the proposed approach for a fine-grained sketch-based image retrieval task on two recently released fine-grained image-sketch datasets [12].

4.1 Experimental Settings

Datasets: In the experiment, two newly released fine-grained SBIR dataset [12] for shoe and chair are utilized. Specifically, there are 419 pairs of photo-sketch samples in the shoe dataset, and 297 pairs of photo-sketch instances in the chair dataset. Attribute annotations are also available for both categories. Taking shoe for example, each shoe is divided into several parts, i.e., toe cap, body, vamp, hell, etc. For each shoe part, a list of part-specific binary attributes are defined. For example, the $1st$ dimension of shoe attribute denotes whether the toe cap is round or not. For a full list of attributes, please refer to [12] instead. It however worth noting that although visual attributes are shared semantic concepts (i.e., toe cap, shoe heel, chair arm, etc.), corresponding photo and sketch attributes for the same shoe do not necessarily agree. This is due to (i) attribute annotations for photos and sketches were conducted independently, and (ii) sketches are often too abstract and iconic to vividly depict certain attributes.

Features: HOG and fc7 Deep [12] are served as features in our experiments. The dimension of HOG is reduced to 210 and 160 for shoe and chair via Principal Component Analysis (PCA), respectively. fc7 Deep is obtained by using the well trained modal provided by [12]. We ran the FG-SBIR experiments for 30 times, and for each time we randomly selected 304/200 pairs of shoe/chair samples for training and took the rest samples for testing.

Evaluation Metric: We follow the same metric used in [21] and [12] for evaluation, i.e., given a query sketch, "*acc.@K*", which is the percentage of relevant photos ranked in the top K results offered by our proposed method.

4.2 Influence of Visual Attributes

To investigate the effect of visual attributes on retrieval result, we choose different sets of attributes as labels for training. More specifically, (i) we divide shoe/chair datasets into three/six subclasses, respectively, (ii) we then select $10d$, $15d$, $21d$ from the original shoe attribute to form new supervision labels; for the chair dataset, the selected dimensions are $5d$, $10d$, and $15d$, and finally (iii) we evaluate the retrieval performances on instance-level. Here, two-view feature via concatenating HOG and $fc7$ deep features is used.

Experiments on each setting are repeated for 30 times, where training and testing data are selected randomly each time. The average retrieval results are reported in Table 1 and Table 2, where we provide retrieval accuracies of @ $K = 1,\ 5,\ 10$. Corresponding plots are also provided in the Fig. 4.

From results on the shoe dataset, we can observe that accuracy on subclass labels is the lowest as expected. The reason is that the subclass labels are a coarse semantic concept and they can not sufficiently capture discriminative visual cues. Furthermore, we discover that attributes with varying dimensions influence the retrieval results dramatically: the more attributes used, the better the results. However, for results on chair (Table 2 and Fig. 4(b)), it is observed that the performance of $5d$ attribute is worse than that of subclass label. The reason is two-fold: (i) the chair attributes introduced by [12] are not overly discriminative (as we also conclude later in Section 4.3), and (ii) the dimensionality of $5d$ is too low to form a discriminative feature representation.

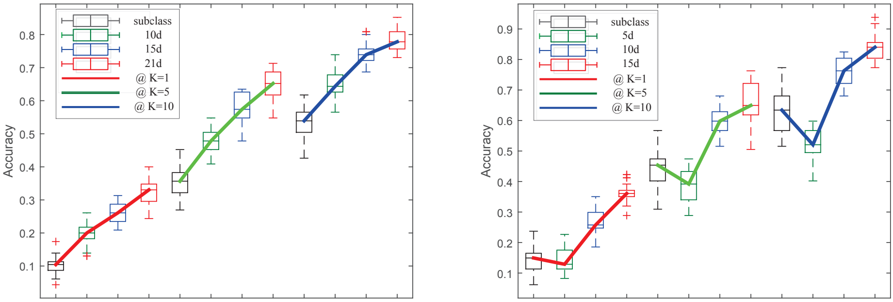
In summary, we can conclude that: (i) attribute labels can be effectively used as supervision information in FG-SBIR; (ii) the dimensionality of the attribute is strongly connected to the capacity of the fine-grained space and has clear effect on retrieval accuracy.

Table 1. Instance-level retrieval accuracies using various attributes on the shoe dataset.

	subclass label	$10d$ attribute	15d attribute	21d attribute
@ $K = 1$	10.23%	19.71%	26.12%	34.78%
@ $K = 5$	35.65%	46.06%	57.74%	64.49%
@ $K = 10$	53.07%	65.30%	74.20%	79.41%

Table 2. Instance-level retrieval accuracies using various attributes on the chair dataset.

	subclass label	$5d$ attribute	$10d$ attribute	$15d$ attribute
@ $K = 1$	14.78%	14.12%	27.04%	36.40%
@ $K = 5$	44.57%	38.42%	59.76%	66.01%
@ $K = 10$	63.78%	52.37%	75.81%	84.54%



(a) Instance-level retrieval accuracies using various attributes on the shoe dataset

(b) Instance-level retrieval accuracies using various attributes on the chair dataset

Fig. 4. Instance-level accuracies. The bold lines colored red, green, and blue denote the retrieval accuracies @ $K = 1,\ 5,\ 10$ respectively. In (a), the boxes colored black, green, blue, and red denote the results obtained by different supervision labels: subclass, $10d$ attribute, $15d$ attribute, $21d$ attribute. In (b), the boxes colored black, green, blue, and red denote the results obtained by different supervision labels: subclass, $5d$ attribute, $10d$ attribute, $15d$ attribute. For each box, the central mark is the median. The top and bottom edges of the box are the 75th and 25th percentiles, respectively. The outliers are marked individually.

4.3 Results of FG-SBIR

Competitors: We mainly benchmark against the very recent deep triplet model proposed in [12]. In addition, we also introduce two shallow variants of our model for comparison:

Deep triplet-ranking: Representing current state-of-the-art for FB-SBIR, the authors [12] develop a deep triplet ranking network with a data augmentation and staged pre-training strategy to address the problem of insufficient training data. We use it for comparison on both the shoe and chair dataset.

$\mathbf{A}_\mathbf{s}$ model: In Section 3, we have illustrated that sketch sample attribute matrix $\mathbf{A}_s$ is usually excessively sparse and low-rank. This is likely to lead to inaccurate computation results, and exactly optimizing for $\mathbf{A}_s\mathbf{T}_s$ that approximate $\mathbf{A}_p$ might not be feasible. In order to verify this, we design the following model for verification and comparison:

$$\begin{aligned} J_2 = \min_{\mathbf{W}_p,\mathbf{W}_s,\mathbf{T}_s} & \|\mathbf{P}^T\mathbf{W}_p - \mathbf{A}_p\|_F^2 + \|\mathbf{S}^T\mathbf{W}_s - \mathbf{A}_s\mathbf{T}_s\|_F^2 \\ & + \lambda_1(\|\mathbf{W}_p\|_{G_1} + \|\mathbf{W}_s\|_{G_1}) + \lambda_2\|\mathbf{A}_p - \mathbf{A}_s\mathbf{T}_s\|_F^2 \ , \end{aligned} \tag{14}$$

where $\mathbf{T}_s$ is the transformation matrices for sketch sample attribute matrix $\mathbf{A}_s$. In the following experiments, we denote this method as "A_s model".

F model: In order to verify the benefits of multi-view features, we introduce models using single-view features for comparison. In Eq. (6), the physical significance of the Group norm terms is view selection. If we set the coefficients of Group norms in Eq. (6) as zero when we use single-view features, and the projection matrices $\mathbf{W}_p$ and $\mathbf{W}_s$ will lose all the constraints. In this case, our model can be adjusted as:

$$\begin{aligned} J_3 = \min_{\mathbf{W}_p,\mathbf{W}_s,\mathbf{T}} & \|\mathbf{P}^T\mathbf{W}_p - \mathbf{A}_p\|_F^2 + \|\mathbf{S}^T\mathbf{W}_s - \mathbf{A}_p\mathbf{T}\|_F^2 \\ & + \lambda_1(\|\mathbf{W}_p\|_F + \|\mathbf{W}_s\|_F) + \lambda_2\|\mathbf{A}_p - \mathbf{A}_p\mathbf{T}\|_F^2 \ . \end{aligned} \tag{15}$$

In the following experiments, we denote this method as "F model". A_s model and F model qualify for shallow model baselines, which are derived from some state-of-the-art shallow cross-modal subspace learning methods elaborated for image-text matching.

Results and Discussion: Results are shown in Table 3. Overall, it can be observed that, on the shoe dataset, our model using concatenation of HOG and $fc7$ deep feature offers the best among all the shallow variants and closely resembles the performance of deep triplet-ranking [12], i.e. 34.78% vs 39.13% for top 1 and 84.54% vs 87.83% for top 10. It is promising to notice that shallow cross-modal method tailored for FG-SBIR is able to deliver retrieval performances close to that of deep models where ample training data and extensive user annotations are required. However, on chairs, our model performed considerably worse than [12], scoring only 36.40% vs 69.07% for top 1 and 84.54% vs 97.04% for top 10. This phenomenon is largely explained by the lack of discriminative power of chair attributes, which was also highlighted as part of previous set of experiments (Section 4.2). We believe redesigning a better set of attributes for chairs would help to boost retrieval performance, but would leave as future work.

In addition, results also show that our model is better than using the single-view feature by "F model", i.e. F model (HOG) and F model (fc7 Deep), and deep feature fc7 Deep is proven to be better than HOG on the FG-SBIR task. It is interesting that when CCA is applied to fuse HOG and fc7 Deep, i.e. F model (HOG&fc7 Deep+2View-CCA), it leads to worse performance when compared against single-view models. The reason is that CCA might result in information loss when fusing features from different modalities. In contrast, our model is capable of keep the properties of the original multi-view features as much as possible via joint view selection. Moreover, in Table 3, we can observe that the experimental results of "A_s model (HOG&fc7 Deep)" on the shoe and the chair datasets are much worse than "Our model (HOG&fc7 Deep)". It indicates that it is more reasonable to use the photo attribute space as the coupled intermediate space for both photo and sketch. In other words, sketch attribute space might suffer from data sparsity and low-rank of attribute matrix $\mathbf{A}_s$, which leads to inefficiency of the model.

Computational Complexity: Average running time of our Matlab code on a 3.30GHz Desktop PC with 16GB RAM, across 30 experiments conducted on the shoe/chair datasets, are 0.87 seconds and 0.39 seconds, respectively.

Table 3. Experimental results comparisons.

	Shoe		Chair	
	acc.@1	acc.@10	acc.@1	acc.@10
Deep triplet-ranking(fc7 Deep) [12]	39.13%	87.83%	69.07%	97.04%
F model (HOG)	3.04%	31.33%	7.22%	42.92%
F model (fc7 Deep)	30.43%	77.91%	35.77%	80.98%
F model (HOG&fc7 Deep+2View-CCA)	6.96%	28.19%	29.00%	70.15%
A_s model (HOG&fc7 Deep)	7.48%	56.52%	24.99%	74.64%
Our model (HOG&fc7 Deep)	34.78%	79.41%	36.40%	84.54%

5 Conclusion

In this paper, for the first time, we proposed an unified cross-domain framework for fine-grained sketch-based image retrieval. Our model not only learns a domain-independent subspace to conduct retrieval, but also ensures effective fine-grained comparisons at the same time. Different to traditional text-photo cross-domain methods that works only on category-level, it uniquely learns from pair-wise sketch-photo data, therefore constructing a coupled space that is fitting for fine-grained retrieval. Once learned the model can also be used to predict attributes without the need for explicit training of attribute classifiers. Experiments on the latest fine-grained sketch-photo datasets demonstrated the effectiveness of the proposed method. For future work, we will investigate how the design of visual attributes affects quality of the learned coupled subspace, with the immediate hope to further improve retrieval performance on the chair dataset.

Acknowledgment

This work was partly supported by National Natural Science Foundation of China (NSFC) grant No. 61402047, Beijing Natural Science Foundation (B-NSF) grant No. 4162044, NSFC-RS joint funding grant Nos. 61511130081 and IE141387, the Open Projects Program of National Laboratory of Pattern Recognition grant No. 201600018, NSFC grant No. 61273217, and the Fundamental Research Funds for the Central Universities under Grant No. $2016RC11$.

References

1. Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., Zhang, L.: Mindfinder: interactive sketch-based image search on millions of images. In: ACM MM, 2010.
2. Cao, Y., Wang, C., Zhang, L., Zhang, L.: Edgel index for large-scale sketch-based image search. In: CVPR, 2011.
3. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE Transactions on Visualization and Computer Graphics, 17(11): 1624–1636, 2011.
4. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. CVIU, 117(7): 790–806, 2013.
5. Lin, Y.L., Huang, C.Y., Wang, H.J., Hsu, W.C.: 3d sub-query expansion for improving sketch-based multi-view image retrieval. In: ICCV, 2013.
6. Parui, S., Mittal, A.: Similarity-invariant sketch-based image retrieval in large databases. In: ECCV, 2014
7. Wang, F., Kang, L., Li, Y.: Sketch-based 3d shape retrieval using convolutional neural networks. In: CVPR, 2015.
8. Ouyang, S., Hospedales, T., Song, Y.Z., Li, X.: Cross-modal face matching: beyond viewed sketches. In: ACCV, 2014.
9. Li, Y., Hospedales, T.M., Song, Y.Z., Gong, S.: Free-hand sketch recognition by multi-kernel feature learning. CVIU, 137: 1–11, 2015.
10. Qi, Y., Guo, J., Song, Y.Z., Xiang, T., Zhang, H., Tan, Z.H.: Im2sketch: Sketch generation by unconflicted perceptual grouping. Neurocomputing, 165: 338–349, 2015.
11. Li, Y., Hospedales, T.M., Song, Y.Z., Gong, S.: Fine-grained sketch-based image retrieval by matching deformable part models. In: BMVC, 2014.
12. Yu, Q., Liu, F., Song, Y., Xiang, T., Hospedales, T., Loy, C.C.: Sketch me that shoe. In: CVPR, 2016.
13. Yu, Q., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-net that beats humans. In: BMVC, 2015.
14. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: ICCV, 2013.
15. Putthividhy, D., Attias, H.T., Nagarajan, S.S.: Topic regression multi-modal latent dirichlet allocation for image annotation. In: CVPR, 2010.
16. Jia, Y., Salzmann, M., Darrell, T.: Learning cross-modality similarity for multinomial data. In: ICCV, 2011.
17. Wu, W., Xu, J., Li, H.: Learning similarity function between objects in heterogeneous spaces. Microsoft Research Technique Report, 2010.
18. Mignon, A., Jurie, F.: Cmml: a new metric learning approach for cross modal matching. In: ACCV, 2012.
19. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation, 16(12): 2639–2664, 2004.
20. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. TPAMI, 29(6): 1005–1018, 2007.
21. Li, K., Pang, K., Song, Y., Hospedales, T., Zhang, H., Hu, Y.: Fine-grained sketch-based image retrieval: The role of part-aware attributes. In: WACV, 2016.
22. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM MM, 2010.

23. Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI, 36(3): 521–535, 2014.
24. Sharma, A., Jacobs, D.W.: Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In: CVPR, 2011.
25. Sharma, A., Kumar, A., Daume III, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: CVPR, 2012.
26. Zhuang, Y., Wang, Y., Wu, F., Zhang, Y., Lu, W.: Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: AAAI, 2013.
27. He, R., Zhang, M., Wang, L., Ji, Y., Yin, Q.: Cross-modal subspace learning via pairwise constraints. TIP, 24(12): 5543–5556, 2015.
28. Rasiwasia, N., Moreno, P.J., Vasconcelos, N.: Bridging the gap: Query by semantic example. TMM, 9(5): 923–938, 2007.
29. Wang, H., Nie, F., Huang, H.: Multi-view clustering and feature learning via structured sparsity. In: ICML, 2013.
30. Gorodnitsky, I.F., Rao, B.D.: Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. IEEE Transactions on Signal Processing, 45(3): 600–616, 1997.