Towards Unsupervised Sketch-based Image Retrieval

Towards Unsupervised Sketch-based Image Retrieval

Conghui Hu[1]
conghui@nus.edu.sg
Yongxin Yang[2]
yongxin.yang@ed.ac.uk
Yunpeng Li[3]
yunpeng.li@surrey.ac.uk
Timothy M. Hospedales[2]
t.hospedales@ed.ac.uk
Yi-Zhe Song[3]
y.song@surrey.ac.uk

[1] National University of Singapore
Singapore
[2] University of Edinburgh
United Kingdom
[3] University of Surrey
United Kingdom

Abstract

The practical value of existing supervised sketch-based image retrieval (SBIR) algorithms is largely limited by the requirement for intensive data collection and labeling. In this paper, we present the first attempt at unsupervised SBIR to remove the labeling cost (both category annotations and sketch-photo pairings) that is conventionally needed for training. Existing single-domain unsupervised representation learning methods perform poorly in this application, due to the unique cross-domain (sketch and photo) nature of the problem. We therefore introduce a novel framework that simultaneously performs sketch-photo domain alignment and semantic-aware representation learning. Technically this is underpinned by introducing joint distribution optimal transport (JDOT) to align data from different domains, which we extend with trainable cluster prototypes and feature memory banks to further improve scalability and efficacy. Extensive experiments show that our framework achieves excellent performance in the new unsupervised setting, and performs comparably to existing zero-shot SBIR methods.

1 Introduction

Sketches efficiently convey the shape, pose and fine-grained details of objects, and thus are particularly valuable in serving as queries to conduct retrieval of photos, *i.e.*, sketch-based image retrieval [32, 41] (SBIR). SBIR has been increasingly well studied, leading to continual improvements in retrieval performance [1, 34]. However state-of-the-art methods generally bridge the sketch-photo domain gap through supervised learning using sketch-photo pairs and class annotation [32, 41]. This supervised learning paradigm imposes a severe bottleneck on the feasibility of SBIR in practice. One main research direction on reducing annotation cost thus far has been zero-shot (category generalized) SBIR [11, 25, 37], where labeled data is no longer necessitated for unseen categories, yet the problem

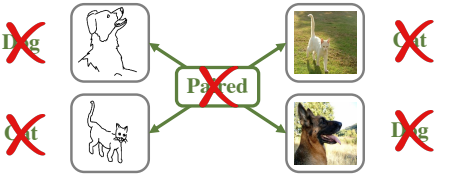

Figure 1: Illustration of unsupervised SBIR where no class label or pairing information is available during training.

still requires availability of all category labels and specific pairing annotations for the *seen* categories. Furthermore, [29] turns images into edge maps to directly mitigate the domain gap, but automatically generated pairs from 3D models are still prerequisites to facilitate effective SBIR.

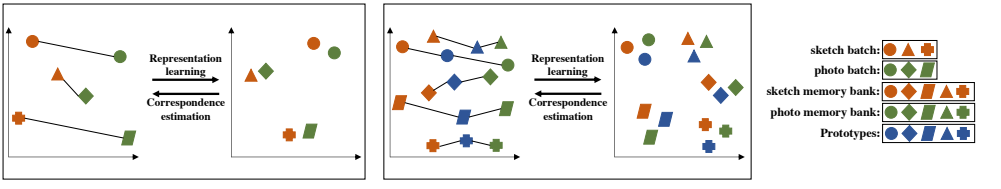

Figure 2: Comparison between batch-wise DeepJDOT [9] and our PM-JDOT. Shapes represent samples of different classes. There are five classes in total for demonstration purposes. (a) In batch-wise DeepJDOT, a single batch only contains samples from a subset of classes, so correspondence is necessarily inaccurate (mismatched shapes/categories linked) and poor alignment is learned. (b) In our PM-JDOT, (i) correspondence is mediated by learned prototypes (blue) for all classes, which enables accurate and efficient computation compactly summarizing the whole distribution; (ii) use of a memory bank allows larger sample size with more unique categories than a single batch, increasing the chance that accurate correspondence can be discovered. Note that hard pairwise correspondence is shown for ease of visualization, but actual OT correspondence computation is soft many-many.

In this paper we go to the extreme in addressing the annotation bottleneck, and study for the first time the problem of *unsupervised* category-level SBIR, where we work under the stringent assumption of (i) no sketch-photo pairing, and (ii) no category annotations (as illustrated in Figure 1) to retrieve photos of the same category as input sketch. We are largely inspired by the recent rapid progress in unsupervised representation learning for photo recognition [5, 6]. However these methods are unsuited to SBIR for the key reason that they are designed for single-domain (photo) representation learning, while SBIR involves *cross-domain* data with a mixture of realistic photos and abstract/iconic sketches. Successful category-level SBIR has thus far relied on sketch-photo pairings and category annotations to drive explicit sketch-photo domain alignment and class-discriminative feature learning prior to retrieval [12, 24, 25, 32, 41]. The key question for us is how such alignment and representation learning can be induced just by working with raw *unpaired and unannotated* photos and sketches.

At a high-level our solution is based on alternating optimization between: (i) computing a soft (many-to-many) correspondence between sketch and photo domains; and (ii) learning

a representation that aligns sketch and photo features under the soft correspondence, and is also semantically meaningful. Our framework is significantly more performant and resistant to local minima compared to hard noisy pairing methods in other applications [15, 44] due to soft correspondence prediction; and a multi-task representation learning objective that synergistically combines cross-domain alignment and in-domain self-supervision for domain-agnostic and semantic-aware feature learning.

In more detail, we first introduce a novel cluster Prototype and feature Memory bank-enhanced Joint Distribution Optimal Transport (PM-JDOT) algorithm for accurate soft cross-domain correspondence estimation. The vanilla JDOT learns to predict cross-domain correspondence using distribution-level information by OT [36]. However, the application of vanilla JDOT in CNNs [9] suffers from an inability to simultaneously provide efficiency and accuracy: OT correspondence is either inaccurate if computed at minibatch level (*e.g.*, a given sketch+photo minibatch likely contains a disjoint set of categories, and thus cannot be correctly aligned, as illustrated in Figure 2(a)); or intractable if computed at dataset level due to $O(N^2)$ cost. We elegantly solve both of these problems by computing OT between cluster prototypes and instances in feature memory bank, which provide a sparse representation of the full dataset; and extending JDOT with features in memory bank to aggregate information across batches (Figure 2(b)). To capture domain-invariant yet class-discriminative features for effective SBIR, we devise an alignment loss to minimize the cross-domain feature discrepancy according to the predicted soft sketch-photo correspondence, and employ a self-supervised loss that helps encode discriminative semantic features by preserving the consistency in cluster assignments between different variants of the same input.

Our main contributions are summarized as follows: (i) We provide the first study of unsupervised SBIR. (ii) We propose a novel unsupervised learning algorithm for multi-domain data that jointly performs cross-domain alignment and semantic-aware feature encoding. (iii) The cluster prototypes and feature memory banks introduced by our PM-JDOT algorithm alleviates the limits of existing JDOT, enabling effective yet tractable distribution alignment. (iv) Extensive experiments on Sketchy-Extended and TUBerlin-Extended datasets illustrate the promise of our framework in both unsupervised and zero-shot SBIR settings.

2 Related Work

Sketch-based image retrieval SBIR methods can be classified into two groups according to granularity: Category-level SBIR aims to rank photos so that those with the same semantic class as the input sketch appear first. Fine-grained SBIR targets on retrieving the specific photo corresponding to the query instance. Traditional supervised SBIR algorithms learn class-discriminative feature using classification loss [32] and remedy the domain gap with sketch-photo paired data [7, 8, 34, 41]. On account of the data shortage that results from labour-intensive sketch-photo paired dataset collection and annotation, zero-shot SBIR intends to test on novel categories that are unseen during training. Representative approaches use adversarial training strategy [12, 24] or triplet ranking loss [31, 40] to learn a common feature space for both domains. Additional side information like word embeddings [11] may also be exploited to preserve semantic information. Nevertheless, annotated training data is still necessary in existing zero-shot SBIR approaches to perform effective training, and the required cross-category generalization is still an active research question [25]. We are therefore motivated to study unsupervised category-level SBIR that does not rely on sketch-photo annotations.

Unsupervised deep learning Unsupervised deep learning methods have recently made strong progress in representation learning that ultimately diminishes the demand for data annotation. Most contemporary unsupervised learning methods can be classified into four categories according to the learning objective: (i) Deep clustering approaches are designed to model the feature space via data grouping where the pseudo class label can be assigned with the help of clustering algorithm [4, 16]. (ii) Instance discrimination [38] treats every single sample as a unique class, which can be beneficial to capture discriminative features of individual instance. (iii) Self-supervised learning algorithms learn through solving different pretext tasks including image colorization [43], image super-resolution [21], image in-painting [27], solving jigsaw puzzle [23], rotation prediction [17]. (iv) Contrastive learning aims to maximize agreement between different augmentations of the same input in feature space [6] or label space [5]. However, these methods are designed for single domain representation learning, and perform poorly if applied directly to multi-domain data. A few self-supervised methods have been defined for multi-domain data [35], but these normally assume that cross-domain pairing is the 'free' pre-text task label, which is exactly the annotation we want to avoid. In contrast, our model performs unsupervised learning in each domain, while simultaneously aligning the domains through JDOT.

Joint distribution optimal transport Optimal transport (OT) [36] is a mathematical theory that enables distance measurement between distributions by way of searching for the optimal transportation plan to match samples from both distributions. OT has been applied in domain adaptation [7, 28, 39] to learn a transportation plan between source and target domains, followed by training a classifier for target domain with transported source domain data and the corresponding category annotation. To avoid this two-step process (feature transformation and classification model training), JDOT [8] aligns the feature-label joint distribution and projects input samples from both domains onto a common feature space where a classifier can be shared. DeepJDOT [9] extends JDOT to deep learning and facilitates training on large scale datasets by introducing a stochastic approximation via batch-wise OT. However, we observe that data in a single batch is not informative enough to represent the whole data distribution, which limits the efficacy of OT in DeepJDOT. To this end, we introduce PM-JDOT which employs prototypes and feature memory banks to enhance representation of each distribution for optimizing OT-based cross-domain alignment.

3 Methodology

In category-level SBIR, the goal is to train an effective CNN $f_\theta : I \rightarrow \mathbf{x}$ to project input imagery I from both sketch and photo domains into a shared embedding space, where features $\mathbf{x}$ facilitate cross-domain instance similarity measurement. Given a query sketch I^s, a ranked list of photos will be generated according to their feature space distance to the query with the aim of ranking photos of the same category on top of the list. In the proposed unsupervised setting, we only have access to a set of training sketches $\mathcal{I}^s = \{I_i^s\}_{i=1}^M$ and photos $\mathcal{I}^p = \left\{I_j^p\right\}_{j=1}^N$ that contain the same categories, but without category or sketch-photo pairing annotations – thus raising the challenge of how to learn a representation suitable for retrieval.

To solve this problem, our method integrates two objectives: (i) cross-domain correspondence estimation with PM-JDOT, which employs the aggregated data in trainable cluster prototypes and feature memory banks in support of accurate and scalable discrepancy measurement. and (ii) unsupervised feature representation learning that encodes domain-agnostic

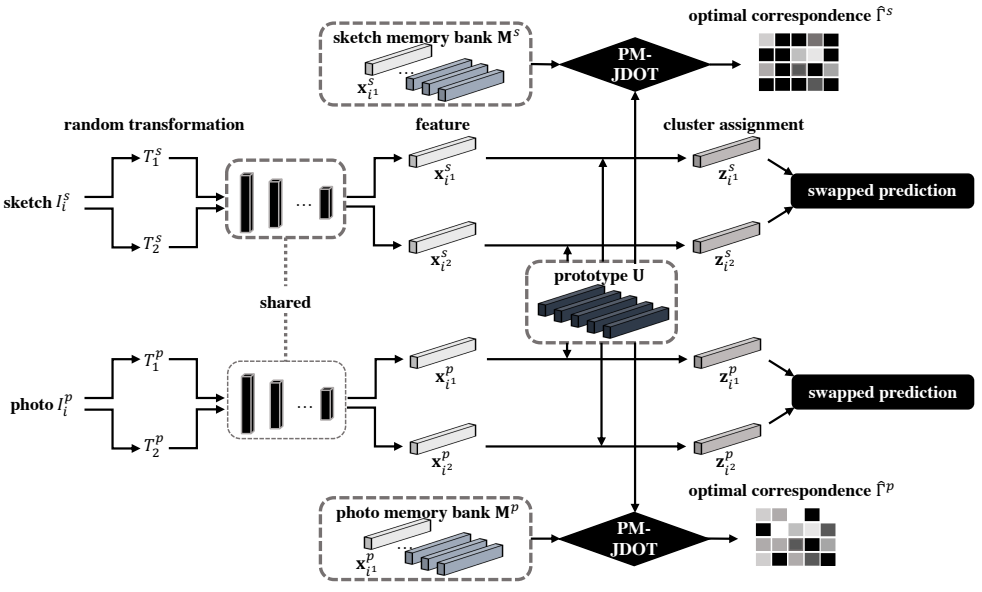

Figure 3: Schematic of our proposed framework.

and semantic-discriminative features from visual input. Figures 3 briefly summarizes our unsupervised SBIR framework.

3.1 Cross-domain correspondence estimation

From JDOT to PM-JDOT Given only unpaired and unlabeled photos and sketches, we introduce a machinery to estimate the soft sketch-photo correspondence in an unsupervised way to support the cross-domain alignment. We introduce joint distribution optimal transport (JDOT) to match samples from the sketch and photo domains. Crucially, we extend it to improve both alignment accuracy and efficiency by redefining the problem in terms of OT between a set of learnable prototypes and feature memory banks – PM-JDOT. Conventional JDOT is able to align all features $\{\mathbf{x}_i^s\}_{i=1}^M$ from sketch domain and $\left\{\mathbf{x}_j^p\right\}_{j=1}^N$ from photo domain via computing the optimal transport plan between them. To make this quadratic computation scale to neural network training, JDOT is applied between two randomly selected batches [9]. However, an individual sketch/photo batch is a weak representation for the overall data distribution of one domain, leading to poor correspondence as illustrated in Figure 2.

Thus we first exploit K trainable cluster prototypes $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K]$ as a stronger proxy to learn a better alignment. Specifically, instead of matching sketch and photo batches in isolation, we estimate the correspondence between sketch/photo batches and the prototypes which compactly represent the whole dataset with a small number of elements. To further alleviate the limitation caused by the impoverished domain representation, *i.e.*, a small batch of samples, we introduce feature memory banks of size E for sketch $\mathbf{M}^s = [\mathbf{x}_1^s, \mathbf{x}_2^s, ..., \mathbf{x}_E^s]$ and photo $\mathbf{M}^p = [\mathbf{x}_1^p, \mathbf{x}_2^p, ..., \mathbf{x}_E^p]$ as richer domain representations which augment current batch with samples in previous batches. The update strategy of memory bank is

FIFO, *i.e.*, removing the oldest batch and putting the current batch on the top of the container.

Correspondence search In PM-JDOT, the correspondence Γ is found from the set of transportation plans Π between prototypes and one feature memory bank by minimizing:

$$\begin{aligned} &\min_{\Gamma\in\Pi}\sum_{i=1}^{K}\sum_{j=1}^{E}\Gamma_{ij}\mathbf{C}(i,j)-\lambda H(\Gamma), \quad \text{where} \\ &\Pi=\left\{\Gamma\in\mathbb{R}_{+}^{K\times E}|\Gamma\mathbf{1}_E=\frac{1}{K}\mathbf{1}_K,\Gamma^{\top}\mathbf{1}_K=\frac{1}{E}\mathbf{1}_E\right\} \end{aligned} \tag{1}$$

Here, $H(\cdot)$ is an entropy regularization term weighted by λ. K and E are the number of prototypes and the size of feature memory bank respectively. The constraint for transportation plans Π ensures each prototype can be selected $\frac{E}{K}$ times on average [5]. $\mathbf{C}\in\mathbb{R}^{K\times E}$ is the matrix of cross-domain pairwise costs. Specifically, the cost $\mathbf{C}(i,j)$ of aligning i^{th} prototype and j^{th} sample in memory bank is calculated by:

$$\begin{aligned} &\mathbf{C}(i,j)=\alpha d_f(\mathbf{u}_i,\mathbf{x}_j)+\beta d_l(\mathbf{v}_i,\mathbf{y}_j), \quad \text{where} \\ &\mathbf{y}_j^{(k)}=\frac{\exp(\mathbf{x}_j^{\top}\mathbf{u}_k/\tau)}{\sum_{m=1}^{K}\exp(\mathbf{x}_j^{\top}\mathbf{u}_m/\tau)} \end{aligned} \tag{2}$$

Here, cosine distance d_f is used to measure the feature-wise similarity between prototype vector $\mathbf{u}_i$ and feature $\mathbf{x}_j$ extracted with f_θ. d_l is applied label-wise to evaluate the difference between one-hot label $\mathbf{v}_i$ for i^{th} prototype and cluster probability $\mathbf{y}_j$ for j^{th} image in the memory bank. $\mathbf{v}_i$ is generated automatically according to the index i, *e.g.*, $\mathbf{v}_1=[0,1,0,0,...,0]$. α and β are scalar hyperparameters that control the contributions of feature and label distance measurements. PM-JDOT is executed twice for prototype-sketch and prototype-photo correspondence, producing optimal correspondence $\hat{\Gamma}^s$ and $\hat{\Gamma}^p$ respectively. Feature extractor f_θ and prototypes $\mathbf{U}$ are fixed in this process.

3.2 Unsupervised representation learning

The algorithm so far in Section 3.1 learns the correspondence between prototypes and samples in both domains, but the feature extractor is not optimized. Thus, in this section, we further illustrate the second part of our alternating optimization: unsupervised representation learning to train f_θ to extract features that are domain-invariant (aligned across domains), yet sensitive to semantic category.

Cross-domain alignment In order to align features from sketch and photo domains, we leverage the first A columns in $\hat{\Gamma}^s$ and $\hat{\Gamma}^p$, which contain the mapping between trainable prototypes and current sketch/photo batch of size A. Then the feature extractor f_θ and trainable prototypes $\mathbf{U}$ are updated by minimizing feature and label discrepancy between corresponding prototypes and samples in the batch according to the optimal correspondences:

$$\begin{aligned} L_a=&L_a^s+L_a^p \\ =&(\sum_{i=1}^{K}\sum_{j=1}^{A}\hat{\Gamma}_{ij}^s(\alpha d_f(\mathbf{u}_i,\mathbf{x}_j^s)+\beta d_l(\mathbf{v}_i,\mathbf{y}_j^s))) \\ &+(\sum_{i=1}^{K}\sum_{j=1}^{A}\hat{\Gamma}_{ij}^p(\alpha d_f(\mathbf{u}_i,\mathbf{x}_j^p)+\beta d_l(\mathbf{v}_i,\mathbf{y}_j^p))) \end{aligned} \tag{3}$$

Semantic-aware feature learning To learn a semantically meaningful representation (*i.e.*, ensure samples from same category are similar in feature space) from unannotated pixel-level input images, inspired by SwAV [6], we train feature extractor f_θ by contrasting the cluster assignments for different variants of the same image. The training objective is to minimize the semantic representation loss:

$$\begin{aligned} L_{se} &= L^s_{se} + L^p_{se} \\ &= (\ell(\mathbf{y}^s_{i1}, \mathbf{z}^s_{i2}) + \ell(\mathbf{y}^s_{i2}, \mathbf{z}^s_{i1})) + (\ell(\mathbf{y}^p_{i1}, \mathbf{z}^p_{i2}) + \ell(\mathbf{y}^p_{i2}, \mathbf{z}^p_{i1})) \end{aligned} \tag{4}$$

where ℓ is the cross-entropy loss. Taking sketch domain for illustration, $\mathbf{y}^s_{it}$ and $\mathbf{z}^s_{it}$ are predicted cluster probability and cluster assignment of I^s_{it} respectively. $\{\mathbf{y}^s_{i1}, \mathbf{z}^s_{i1}\}$ and $\{\mathbf{y}^s_{i2}, \mathbf{z}^s_{i2}\}$ correspond to two transformed variants $I^s_{i1} = T_1(I^s_i)$ and $I^s_{i2} = T_2(I^s_i)$ of the same original sketch I^s_i, where T_1 and T_2 are randomly sampled from the set $\mathcal{T}$ of image transformations including rescaling, flipping, *etc*. Through *swapped prediction*, *i.e.*, pairing $\mathbf{y}^s_{i1}$ with $\mathbf{z}^s_{i2}$ and $\mathbf{y}^s_{i2}$ with $\mathbf{z}^s_{i1}$ in cross-entropy loss ℓ, the network learns to predict consistent cluster probabilities for different augmentations of identical image, which assists semantically-aware feature learning. $\mathbf{y}^s_{it}$ can be measured in the same way as Equation 2. And we compute cluster assignment $\mathbf{z}^s_{it}$ online at each iteration as follow [6]:

$$\begin{aligned} &\max_{\mathbf{Z}\in\mathcal{Z}} \mathrm{Tr}(\mathbf{Z}^\top \mathbf{U}^\top \mathbf{Q}) + \varepsilon H(\mathbf{Z}), \quad \text{where} \\ &\mathcal{Z} = \left\{ \mathbf{Z} \in \mathbb{R}^{K\times B}_+ | \mathbf{Z}\mathbf{1}_B = \frac{1}{K}\mathbf{1}_K, \mathbf{Z}^\top \mathbf{1}_K = \frac{1}{B}\mathbf{1}_B \right\} \end{aligned} \tag{5}$$

Where $\mathbf{Q}$ is a feature queue of size B which is initialized with image features and updated continuously in a FIFO manner during training. If the training batch size is A, the current batch features define the top A elements in $\mathbf{Q}$. $\mathbf{Z}$ are cluster assignments corresponding to the B samples in $\mathbf{Q}$. $\mathbf{U}$ represents cluster prototypes. $H(\cdot)$ is an entropy penalty with weight ε. Only the cluster assignments for current batch, *i.e.*, top A elements in $\mathbf{Z}$, are used for L_{se}.

Summary The overall learning objective is to train an effective feature extractor f_θ without class or instance-paired annotation. We achieve this by minimizing alignment loss L_a and the semantic representation loss L_{se} as:

$$\underset{\theta,\mathbf{U}}{\operatorname{argmin}}\ \nu L_a + \mu L_{se} \tag{6}$$

where ν and μ are respective loss weights. Algorithm 1 in Supplementary material summarizes the training algorithm followed in this work.

4 Experiments

4.1 Datasets and Settings

Datasets We evaluate our algorithm on two datasets: (i) Sketchy-Extended [22] contains 75,471 free-hand sketches and 12,500 photos spanning over 125 categories provided by [32] and another 60,502 photos collected in [22] from ImageNet [10]. (ii) TUBerlin-Extended [42] offers 20,000 sketches [13] evenly distributed on 250 classes and photos of same categories collected using Google image search.

Implementation details We use ResNet-50 [19] as feature extractor f_θ, followed by an additional L2 normalization layer to transform visual input into 128-d feature embeddings.

Table 1: Unsupervised SBIR results on Sketchy-Extended and TUBerlin-Extended dataset

Method	Sketchy-Extended dataset			TUBerlin-Extended dataset		
	Prec@200(%)	mAP@200 (%)	mAP (%)	Prec@200(%)	mAP@200 (%)	mAP (%)
RotNet [10]	2.26	4.89	1.54	1.53	3.61	0.77
ID [33]	3.41	5.26	2.45	2.66	5.35	1.35
CDS [20]	2.37	3.58	1.88	2.64	4.69	1.63
GAN [18]	2.45	4.66	1.43	1.56	3.45	0.69
SwAV [5]	10.87	12.51	10.15	3.36	5.81	2.89
DSM [28]	10.07	17.92	4.28	7.05	13.00	2.61
SwAV [5] + CycleGAN [43]	4.15	5.39	4.28	2.67	3.50	2.06
SwAV [5] + GAN [18]	22.96	25.48	18.82	10.92	13.46	8.43
Ours	**33.64**	**36.31**	**28.17**	**14.78**	**18.66**	**9.93**

f_θ is first initialized with parameters pre-trained with photos in ImageNet dataset [10] by applying SwAV [5]. As SwAV [5] is an unsupervised learning framework, it is guaranteed that *no labeled data is used in the pre-training stage*. All training photo features extracted with the pre-trained f_θ are grouped into K clusters using K-means. The K cluster centroids are then employed to initialize prototypes $\mathbf{U}$. The number of prototypes K is set to the actual number of training categories, *i.e.*, 125 for Sketchy-Extended and 250 for TUBerlin-Extended in unsupervised SBIR. The sum of elements related to current batch in $\hat{\Gamma}^s$ and $\hat{\Gamma}^p$ are normalized to 1 in Equation 3 for all experiments. Both the feature extractor and prototypes are trained with learning rate initialized with 0.01 and divided by 2 after each 10 epochs. We use SGD optimizer and set momentum factor and weight decay value to 0.9 and 1e-4 respectively. Weights for L_a and L_{se} are 1 and 10. And temperature hyperparameter τ is set to 0.1. Our framework is implemented with Pytorch [26] and optimal transportation plans are computed by the POT toolbox [14].

Evaluation metrics Cross-domain retrieval is performed by computing cosine distance between sketch and photo feature vectors and generating a ranked list of gallery photos. We evaluate the retrieval performance by calculating the precision and mean average precision among top 200 retrieved photos denoted by Prec@200 and mAP@200 as well as the mean average precision over the whole dataset (mAP). Photos belonging to the same category as the query sketch are considered as correct retrievals.

4.2 Results

4.2.1 Unsupervised SBIR

Settings 50 and 10 sketches for each class are randomly selected as query sets for Sketchy-Extended and TUBerlin-Extended dataset respectively for testing. The remaining sketches and photos are used during the training process by following the same setting in [22]. No category labels or sketch-photo pairings are available during training. Each mini-batch contains 128 96×96 pixel sketches and photos.

Results Quantitative retrieval results on Sketchy-Extended and TUBerlin-Extended are shown in Table 1. From the results, we make the following observations: (i) Unsupervised feature representation learning algorithms [5, 17, 33] originally designed for single-domain perform poorly when directly applied to cross-domain task like SBIR. SwAV [5] is the best among these three methods. (ii) CDS [20] cannot cope with the large domain gap between sketch and photo and results in unsatisfactory performance. (iii) From the comparison between GAN [18] and SwAV+GAN, we can see that additional guidance targeting on preserving semantic-discriminative feature is essential in category-level SBIR. (iv) CycleGAN fails to generate high-quality color images from sketch in large-scaled multi-class image trans-

Figure 4: Top8 retrieval results for unsupervised SBIR. Row 1&5: Retrieval results of SwAV [5]; Row 2&6: Retrieval results of DSM [29]; Row 3&7: Retrieval results of SwAV [5] + GAN [18]; Row 4&8: Retrieval results of our framework.

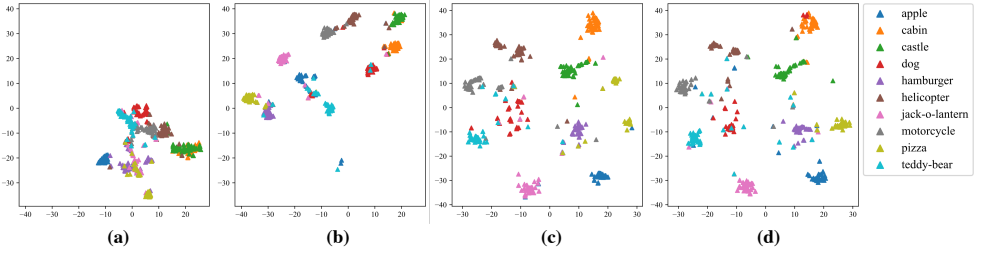

Figure 5: *t-SNE* visualization of 10 categories from Sketchy-Extended dataset. (a): Sketch feature visualization of SwAV [5]; (b): Photo feature visualization of SwAV [5]; (c): Sketch feature visualization of our method; (d) Photo feature visualization for our method.

lation. In contrast, it degrades the semantic information in the original sketch and leads to worse retrieval results compared with SwAV only. (v) Our proposed framework achieves the best retrieval accuracy compared with all these baseline methods trained without external labeled data. Quantitative retrieval results and feature visualizations can be found in Figure 4 and Figure 5.

4.2.2 Zero-shot SBIR

Settings We use the same data split as [11]: 104 and 21 categories are selected for training and testing respectively for Sketchy-Extended dataset. 30 classes are randomly chosen from TUBerlin-Extended dataset for testing and the rest are used for training. Following the default setting in [11], we set each mini-batch to 20 224×224 sketches and photos.

Results Retrieval performance in Table 2 shows that even without involving human pair-

Table 2: Zero-shot SBIR results on Sketchy-Extended and TUBerlin-Extended dataset. $(*)^a$ represents for retrieval results on 25 test categories following the setting proposed in [30]. All methods except ours use instance-wise annotation in the train set.

Method	Supervision	Sketchy-Extended dataset			TUBerlin-Extended dataset		
		Prec@200(%)	mAP@200 (%)	mAP (%)	Prec@200(%)	mAP@200 (%)	mAP (%)
ZSIH [23]	✓	-	-	25.90^a	-	-	23.40
CVAE [11]	✓	33.30	22.50	19.59	0.30	0.90	0.50
SAN [42]	✓	32.20	23.60	-	21.80	14.10	-
SEM-PCYC [3]	✓	-	-	34.90^a	-	-	**29.70**
Doodle [31]	✓	37.04	**46.06**	**36.91**	12.08	15.68	10.94
Ours	✗	**38.44**	44.09	34.68	**28.36**	**31.53**	22.91

Table 3: Ablation study on our model components. Unsupervised SBIR on Sketchy-Extended and TUBerlin-Extended dataset.

Method	JDOT	Proto.	Mem. bank	Sketchy-Extended dataset			TUBerlin-Extended dataset		
				Prec@200(%)	mAP@200 (%)	mAP (%)	Prec@200(%)	mAP@200 (%)	mAP (%)
v1	✗	✗	✗	10.87	12.51	10.15	3.36	5.81	2.89
v2	✓	✗	✗	21.07	23.19	18.53	7.01	9.96	5.05
v3	✓	✓	✗	25.60	28.62	20.98	9.01	12.26	5.62
v4	✓	✗	✓	24.83	27.67	20.78	11.71	15.66	7.53
v5	✓	✓	✓	**33.64**	**36.31**	**28.17**	**14.78**	**18.66**	**9.93**

wise or category-level annotations during training, our framework still performs comparably with existing zero-shot SBIR algorithms that use such annotations during training. Our aligned semantically rich and domain-invariant representation learned on unlabeled training data can generalize directly to unseen classes not used for training.

4.2.3 Ablation Study

We analyze the efficacy of different components in our unsupervised SBIR framework in Table 3: (i) Compared with vanilla SwAV (v1), JDOT using batch-wise OT (v2) for alignment as in [9] already benefits cross-domain matching in both datasets; (ii) In v3, the transportation map is measured between prototypes and single batch of instances. The result shows that prototypes offers a better approximation for real data distribution and improves the OT-based alignment; (iii) Making use of additional data for memory bank-wise OT (v4) is also beneficial for feature alignment; and (iv) Our full model (v5), which takes advantages of both prototypes and memory banks, provides best alignment and representation learning strategy. Further analysis can be found in the Supplementary Material.

5 Conclusion

This paper presents the first attempt at unsupervised SBIR, which is a more challenging learning problem, but more practically valuable due to addressing the data annotation bottleneck. To facilitate cross-domain feature representation learning with no labeled data, our proposed framework performs cross-domain correspondence estimation and unsupervised representation learning alternatively. Alignment is further achieved accurately and scalably by our PM-JDOT. The results show that our unsupervised framework already provides usable performance on par with contemporary zero-shot SBIR methods, but without requiring any instance-wise category or pairing annotation.

References

[1] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9779–9788, 2020.

[2] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 999–1008, 2022.

[3] Ayan Kumar Bhunia, Aneeshan Sain, Parth Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive fine-grained sketch-based image retrieval, 2022. URL https://arxiv.org/abs/2207.01723. arXiv:2207.01723.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 132–149, 2018.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2020. URL http://arxiv.org/abs/2006.09882. arXiv:2006.09882.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. on Mach. Learn. (ICML)*, pages 1597–1607, 2020.

[7] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint Euro. Conf. on Mach. Learn. and Knowl. Discov. in Databases (ECML)*, pages 274–289, 2014.

[8] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation, 2017. URL http://arxiv.org/abs/1705.08848. arXiv:1705.08848.

[9] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 447–463, 2018.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 248–255, 2009.

[11] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2179–2188, 2019.

[12] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5089–5098, 2019.

[13] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graphics*, 31(4):1–10, 2012.

[14] R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017. URL https://pythonot.github.io/.

[15] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 6112–6121, 2019.

[16] Boyan Gao, Yongxin Yang, Henry Gouk, and Timothy M Hospedales. Deep clustering with concrete k-means. In *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pages 4252–4256, 2020.

[17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2020. URL http://arxiv.org/abs/1803.07728. arXiv:1803.07728.

[18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL http://arxiv.org/abs/1406.2661. arXiv:1406.2661.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1105–1113, 2016.

[20] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels, 2020. URL http://arxiv.org/abs/2003.08264. arXiv:2003.08264.

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4681–4690, 2017.

[22] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2862–2871, 2017.

[23] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 69–84, 2016.

[24] Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, Anurag Mittal, and Hema Murthy. Stacked adversarial network for zero-shot sketch based image retrieval. In *Proc. IEEE/CVF Winter Conf. on Appl. of Comput. Vis. (WACV)*, pages 2540–2549, 2020.

[25] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 677–686, 2019.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library, 2019. URL http://arxiv.org/abs/1912.01703. arXiv:1912.01703.

[27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2536–2544, 2016.

[28] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Proc. NeurIPS*, pages 4204–4212, 2016.

[29] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 751–767, 2018.

[30] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *British Mach. Vis. Conf. (BMVC)*, volume 1, page 7, 2015.

[31] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7462–7471, 2022.

[32] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graphics*, 35(4):1–12, 2016.

[33] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3598–3607, 2018.

[34] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 5551–5560, 2017.

[35] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2019. URL http://arxiv.org/abs/1906.05849. arXiv:1906.05849.

[36] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[37] Wenjie Wang, Yufeng Shi, Shiming Chen, Qinmu Peng, Feng Zheng, and Xinge You. Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval. In *Int. Joint Conf. on Artif. Intell. (IJCAI)*, pages 1106–1112, 2021.

[38] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3733–3742, 2018.

[39] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *Int. Joint Conf. on Artif. Intell. (IJCAI)*, volume 7, pages 2969–2975, 2018.

[40] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 300–317, 2018.

[41] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 799–807, 2016.

[42] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016.

[43] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 649–666, 2016.

[44] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 8222–8231, 2019.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 2223–2232, 2017.