Feature Representation Learning for Unsupervised Cross-domain Image Retrieval

Conghui Hu[0000−0002−4984−3960] and Gim Hee Lee[0000−0002−1583−0475]

Department of Computer Science, National University of Singapore
{conghui,gimhee.lee}@nus.edu.sg

Abstract. Current supervised cross-domain image retrieval methods can achieve excellent performance. However, the cost of data collection and labeling imposes an intractable barrier to practical deployment in real applications. In this paper, we investigate the unsupervised cross-domain image retrieval task, where class labels and pairing annotations are no longer a prerequisite for training. This is an extremely challenging task because there is no supervision for both in-domain feature representation learning and cross-domain alignment. We address both challenges by introducing: 1) a new cluster-wise contrastive learning mechanism to help extract class semantic-aware features, and 2) a novel distance-of-distance loss to effectively measure and minimize the domain discrepancy without any external supervision. Experiments on the Office-Home and DomainNet datasets consistently show the superior image retrieval accuracies of our framework over state-of-the-art approaches. Our source code can be found at https://github.com/conghuihu/UCDIR.

Keywords: Unsupervised feature representation learning, Cross-domain alignment

1 Introduction

Cross-domain image retrieval refers to the task where the imagery data in one domain is used as query to retrieve the relevant samples in other domains. This task has many useful applications in our daily life. For example, sketch-based photo retrieval can be used in online shopping to search a product. To facilitate effective cross-domain retrieval, existing works takes the annotated class labels [24] or cross-domain pairing information [31] as supervision to train the model. However, it is always expensive and tedious to annotate labels for both domains, which severely limits the practical value of previous fully-supervised works. This limitation motivates us to circumvent the requirement for large amounts of annotated ground truth data by investigating the task of unsupervised cross-domain image retrieval. We thus focus on the category-level unsupervised cross-domain image retrieval. Specifically, the goal is to train a network to retrieve images from the same category using the query image across a different domain without any annotated class labels and cross-domain pairing information for the training data. Two challenges must be solved to achieve the goal of category-level unsupervised cross-domain image retrieval: 1) effectively bridge the gap between an

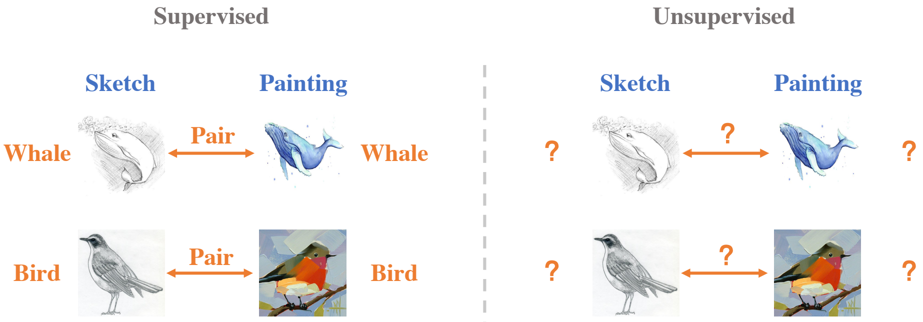

Fig. 1. Illustration of unsupervised cross-domain image retrieval. Compared with its supervised counterpart on the left, class label and pair annotation are not accessible in our unsupervised setting.

imagery input and its corresponding semantic concept without label supervision, and 2) align the data between different domains without any cross-domain pair annotation.

The first challenge we faced is common in unsupervised feature representation learning [30,1,15], whose objective is to extract discriminative feature representation from pixel-level input without class annotations. Nonetheless, unsupervised feature representation learning does not consider domain shifts and thus would fail catastrophically when directly applied to our category-level unsupervised cross-domain image retrieval task due to the domain gap. The second challenge we faced is closely related to the line of works in unsupervised domain adaptation [32], where an unlabeled target domain needs to be aligned with the source domain to accurately classify the data in the target domain. However, fully [4] or partially [32] labeled source domain data are normally available for unsupervised domain adaptation. As a result, the task of unsupervised domain adaptation is easier than our category-level unsupervised cross-domain image retrieval since the labels in the source domain data can be used to learn discriminative features in the source domain and then transferred to the unlabeled target domain. Furthermore, our goal is to retrieve image of same category from the other domain, while unsupervised domain adaptation algorithms mostly focus on image classification.

In this paper, we formulate a novel end-to-end learning framework which incorporates both in-domain unsupervised representation learning and cross-domain alignment to accomplish our objective of training cross-domain image retrieval model in an unsupervised manner. We address the first challenge by introducing a new cluster-wise contrastive learning mechanism to help extract class semantic-aware features. In contrast to existing instance-wise contrastive learning loss [20] which neglects class semantic by only considering the augmented views of itself as positive samples, our cluster-wise contrastive loss is based on feature clusters that pulls samples of similar semantics closer and pushes different clusters apart. We address the second challenge by proposing a novel distance-of-distance loss which is able to effectively measure and minimize the domain discrepancy without external supervision. Specifically, we circumvent the difficulty of domain alignment due to the unknown class labels associated to the feature clusters in each domain by designing our distance-of-distance loss to be invariant to the cluster orders.

Our main contributions can be summarized as the follows:

1. We develop a novel feature representation learning algorithm for unsupervised cross-domain image retrieval.
2. To enable semantic-aware feature extraction, we propose a new cluster-wise contrastive learning loss to minimize the feature distances between semantically similar samples.
3. A novel distance-of-distance loss is carefully designed to measure domain discrepancy and help achieve cross-domain alignment.
4. Extensive experiments on the Office-Home and DomainNet datasets demonstrate the efficacy of our proposed framework.

2 Related Work

2.1 Cross-domain Image Retrieval

Image retrieval is a long-standing research problem in computer vision. Given a query image, the objective is to retrieve images that meet certain predefined criteria from the database [26]. The criteria can be category-level or instance-level correspondence according to the granularity. All images in the database are ranked according to the distance to the query image [18]. Thus, the task boils down to effectively measure the similarity between images. It becomes more challenging when the query and images in database are from different domains, *i.e.*, cross-domain image retrieval. For example, the query can be just a free-hand sketch with a few strokes, while images in the database are all real photographs [24]. To accurately compute the distance for images across domains, class labels are required as the supervision [24] for class semantic-aware feature extraction and cross-domain pairing annotations are leveraged to bridge the domain gap by minimizing triplet [31,33] or HOLEF loss [27]. However, the annotations used as supervision for model training are always labour-intensive to source. We are therefore motivated to investigate the unsupervised cross-domain image retrieval that shares all challenges laid out in cross-domain image retrieval but without external supervision.

2.2 Unsupervised Representation Learning

Unsupervised representation learning has been actively studied in recent years. Deep clustering related methods attempt to assign a pseudo class label for each sample by traditional clustering methods such as K-means [1] or maintain a set of trainable cluster centroids [7], and pseudo labels are continually refined during training. In contrast to deep clustering, instance discrimination [30] directly regards every sample itself as a standalone class and the training objective is to distinguish the sample from all the rest of data for the sake of extracting meaningful discriminative features. Self-supervised learning approaches facilitate the representation learning by introducing various pretext tasks like image rotation prediction [8], jigsaw puzzle solving [19] and image in-painting [22]. Supervisions for all these pretext tasks are free to obtain. Contrastive learning has been increasingly gaining popularity in unsupervised representation learning due to its effectiveness. The goal of contrastive learning algorithms is to maximize the agreement between positive pairs like two augmented views of the same images [2] or the image and its corresponding cluster centroid [15]. Nevertheless, the aforementioned unsupervised representation learning methods are originally devised for single domain data. The large gap between different domains preemptively limits their practical value.

2.3 Unsupervised Domain Adaptation

A related line of research that addresses the domain gap without full-supervision is unsupervised domain adaptation. Conventional unsupervised domain adaptation targets on transferring knowledge learned from an annotated source domain to a novel unlabeled target domain. The model can be trained to predict semantic-aware features for the source-domain data with the help of source domain label. The key challenge then becomes aligning the target domain data with their counterpart in the source domain. The domain discrepancy can be directly measured by Maximum Mean Discrepancy (MMD) [10] or Joint MMD [17], and minimized to remedy the domain gap [16,17]. [25,4] managed to search for matching data pairs across domains through Optimal Transport [29]. Moreover, Generative adversarial Networks [9] can also be utilized to break the domain barrier in either feature-level [6] or pixel-level [12]. More recently, a new domain adaptation setting with only few-shot annotated samples from source domain is introduced to further reduce the data labeling cost. [13,14] employed instance discrimination [30] for both in-domain and cross-domain to learn a shared and instance-discriminative feature space. In [32], prototypes are applied to make the feature more semantic-aware. Our unsupervised cross-domain image retrieval is different from domain adaptation in two aspects: 1) There is no requirement for labeled data. Both source and target domain are unlabeled; 2) Our task is image retrieval rather than image classification.

3 Our Method

3.1 Overview

We use domain A and domain B to denote the domain shift in images. Given a query image I_i^A of category k in domain A, category-level cross-domain image retrieval is considered successful when images of the same category k in domain B are retrieved. To accomplish the goal of cross-domain retrieval, it is required to train a valid feature extractor $f_\theta : I \mapsto \mathbf{x}$ which can project input image I from both domains to feature $\mathbf{x}$ in a common embedding space. All images $\mathcal{I}^B = \left\{I_j^B\right\}_{j=1}^M$ in domain B are then ranked by the feature distance $d(\mathbf{x}_i^A, \mathbf{x}_j^B)$ between I_j^B and the query image I_i^A in domain A. For I_i^B of category k to appear on top of the list, feature extractor f_θ needs to be capable of learning: 1) class semantic-aware features to discriminate samples among different classes; and 2) domain-agnostic features to facilitate the direct distance measurement between images in domain A and B. However, only a set of unlabeled images $\mathcal{I}^A = \left\{I_i^A\right\}_{i=1}^N$ and $\mathcal{I}^B = \left\{I_j^B\right\}_{j=1}^M$ from domain A and B are provided for training under the unsupervised cross-domain image retrieval setting. As a result, learning semantically meaningful and domain-invariant feature embeddings becomes extremely challenging. To learn feature representations that fulfill the aforementioned requirements, we design our framework according to the following two aspects: 1) In-domain representation learning which targets on learning class-discriminative features through a novel cluster-wise contrastive learning mechanism; 2) Cross-domain alignment through the distance of distance minimization as shown in Figure 2.

3.2 In-domain Cluster-wise Contrastive learning

Instance-wise contrastive learning methods take only augmented views of the same instance as positive pair, while all other samples in the dataset are regarded as the negatives. The loss function [20] is defined as:

$$\mathcal{L}_{\text{IW}} = \sum_{i \in \mathbf{I}} - \log \frac{\exp(\mathbf{x}_i^\top \mathbf{x}_i' / \tau)}{\sum\limits_{a \in \mathbf{I}} \exp(\mathbf{x}_i^\top \mathbf{x}_a' / \tau)}, \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_i'$ can be feature embeddings from different augmented views of instance i. $\mathbf{I}$ represents indices of all the samples in the same domain.

Although all instances are well-seperated at instance-level after training the feature extractor with $\mathcal{L}_{\text{IW}}$, they are not clustered together according to their classes. However, the feature space that can encode the class semantic structure of the data is desired in the cross-domain image retrieval task. This motivates us to design the cluster-wise contrastive learning loss. Specifically, there are two steps in our cluster-wise contrastive learning: 1) image clustering and 2) contrastive learning with pseudo label. Additionally, we propose to perform cluster-wise contrastive learning separately in domain A and B. For brevity, we remove the domain notation in this section.

Image clustering. Following MoCo [3], we maintain a momentum encoder $f_{\theta'}$ to extract features for image clustering as it yields more consistent clusters. Let $\mathbf{x}_i$ and $\mathbf{x}_i'$ be the features extracted from the trainable encoder f_θ and momentum encoder $f_{\theta'}$, respectively. We apply K-means on all image features $\{\mathbf{x}_i'\}_{i=1}^N$ in one single domain to obtain its K clusters. Each sample I_i is assigned with a pseudo label y_i according to the K-means results. All pseudo labels are updated after every epoch.

Contrastive learning with pseudo label. Given the pseudo labels predicted by image clustering, we can now conduct cluster-wise contrastive learning. Samples in the same cluster as the query are used to form positive pairs. As a result, the feature extractor is trained to pull feature distances within the same cluster closer, while pushing different clusters apart. Specifically, the learning objective of cluster-wise contrastive learning is given by:

$$\mathcal{L}_{\text{CW}} = \sum_{i \in \mathbf{I}} \frac{-1}{|\mathbf{P}(i)|} \sum_{p \in \mathbf{P}(i)} \log \frac{\exp(\mathbf{x}_i^\top \mathbf{x'}_p / \tau)}{\sum\limits_{a \in \mathbf{I}} \exp(\mathbf{x}_i^\top \mathbf{x'}_a / \tau)}, \tag{2}$$

where $\mathbf{I}$ denotes indices of all the samples in the same domain and $\mathbf{P}(i)$ represents the indices for the set of samples belonging to the same cluster as I_i, *i.e.*, $\mathbf{P}(i) = \{p \in \mathbf{I} : y_p = y_i\}$ and $|\mathbf{P}(i)|$ is its cardinality.

To encourage local smoothness and generate valid clustering results at the beginning of the training process, we also add the instance-wise contrastive loss. Therefore, our loss for in-domain feature representation learning becomes:

$$\mathcal{L}_{\text{in-domain}} = \mathcal{L}_{\text{IW}} + \lambda \mathcal{L}_{\text{CW}}. \tag{3}$$

Since the clustering results are not sufficient reliable at the initial learning stage, we gradually increase the weight for $\mathcal{L}_{CW}$ and set λ as:

$$\lambda = \begin{cases} 0 & ep <= T_1 \\ \alpha \frac{ep - T_1}{T_2 - T_1} & T_1 < ep < T_2 \\ \alpha & ep >= T_2 \end{cases}, \tag{4}$$

where α is a weight hyper-parameter. ep, T_1 and T_2 are the current training epoch, the number of epoch to include $\mathcal{L}_{CW}$, the number of epoch to stop increasing the weight for $\mathcal{L}_{CW}$, respectively.

3.3 Cross-domain Alignment

Domain-invariant is another requirement for the features in cross-domain image retrieval. However, it is difficult to effectively align feature clusters across domains when there is no class label nor correspondence annotation that can be utilized as supervision in our unsupervised setting. Furthermore, the order of the predicted cluster centroids $\left\{\mathbf{c}_u^A\right\}_{u=1}^K$ and $\left\{\mathbf{c}_u^B\right\}_{u=1}^K$ for domain A and B are not fixed since we perform K-means separately in the two domains. The unknown correspondences between the clusters across the two domains increases the challenge for cross-domain alignment. To solve this problem, we propose to measure the cross-domain distance of the in-domain distance. In other words, the Distance-of-Distance (DD) loss.

In-domain Distance. Given an input image I_i, we calculate its clustering probabilities using the cluster centroids, *i.e.*:

$$p_i^u = \frac{\exp(\mathbf{x}_i^\top \mathbf{c}_u / \phi)}{\sum\limits_{k=1}^{K} \exp(\mathbf{x}_i^\top \mathbf{c}_k / \phi)}, \tag{5}$$

where ϕ is a temperature hyper-parameter. $\mathbf{c}_u$ represents the centroids for cluster u. The clustering probabilities for I_i is $\mathbf{p}_i = [p_i^{(1)}, p_i^{(2)} ..., p_i^{(K)}]^\top$. We leverage the centroids for domain A and B to obtain $\mathbf{p}_i^A$ and $\mathbf{p}_i^B$, respectively. The in-domain distance for domain A is defined as:

$$\begin{aligned} &d_{ij}^A = \mathrm{D}(\mathbf{p}_i^A, \mathbf{p}_j^A), \quad \text{where} \\ &\mathbf{p}_i^A = [p_i^{(A^1)}, p_i^{(A^2)} ..., p_i^{(A^K)}]^\top, \quad \mathbf{p}_j^A = [p_j^{(A^1)}, p_j^{(A^2)} ..., p_j^{(A^K)}]^\top. \end{aligned} \tag{6}$$

Here, $\mathrm{D}(\cdot,\cdot)$ is the cosine distance. This in-domain distance measures the difference in clustering probabilities for samples.

Proposition 1. *Value of d_{ij}^A remains the same when the order of the elements in $\mathbf{p}_i^A$ and $\mathbf{p}_j^A$ are simultaneously shuffled, i.e., order of the centroids is changed.*

Proof. Suppose we randomly shuffle corresponding elements of $\mathbf{p}_i^A$ and $\mathbf{p}_j^A$ to:

$$\mathbf{p'}_i^A = [p_i^{(A^K)}, p_i^{(A^1)}, \cdots, p_i^{(A^2)}]^\top \quad \text{and} \quad \mathbf{p'}_j^A = [p_j^{(A^K)}, p_j^{(A^1)}, \cdots, p_j^{(A^2)}]^\top,$$

we get $d'^A_{ij} = \mathrm{D}(\mathbf{p'}_i^A, \mathbf{p'}_j^A) = 1 - \eta'^{-1}(p_i^{(A^K)} p_j^{(A^K)} + p_i^{(A^1)} p_j^{(A^1)} + \cdots + p_i^{(A^2)} p_j^{(A^2)}) = 1 - \eta^{-1}(p_i^{(A^1)} p_j^{(A^1)} + p_i^{(A^2)} p_j^{(A^2)} + \cdots + p_i^{(A^K)} p_j^{(A^K)}) = \mathrm{D}(\mathbf{p}_i^A, \mathbf{p}_j^A) = d_{ij}^A$, where $\eta' = \|\mathbf{p'}_i^A\|\|\mathbf{p'}_j^A\| = \|\mathbf{p}_i^A\|\|\mathbf{p}_j^A\| = \eta$ due to the commutative property of addition. □

Corollary 1. *Our proposed in-domain distance has the important property of order-invariant. This order-invariant property also holds for d_{ij}^B for domain B.*

Cross-domain Distance-of-Distance. As mentioned previously, our key challenge is to design a proper discrepancy measurement method to align the two domains while the cluster orders are unknown. Provided with the order-invariant in-domain distance, we devise a new cross-domain distance of distance:

$$dd_{ij} = \mathrm{DD}(d_{ij}^A, d_{ij}^B). \tag{7}$$

where $\mathrm{DD}(\cdot,\cdot)$ represents the distance-of-distance (DD) calculator that measures the L2 distance between two in-domain distances. As d_{ij}^A and d_{ij}^B are both order-invariant and the data samples used in the in-domain distance calculation are the same I_i and I_j. The value of dd_{ij} is then only related to the difference between values in the two sets of centroids $\{\mathbf{c}_u^A\}_{u=1}^K$ and $\{\mathbf{c}_u^B\}_{u=1}^K$ regardless of the order of cluster centroids. For two well-aligned domains A and B, dd_{ij} is small since the centroids for the two domains are similar. However, the corresponding centroids becomes disparate when there is a big difference between the feature distribution for domain A and B. Consequently, the value of d_{ij}^A and d_{ij}^B would differ greatly and thus lead to a large dd_{ij}. The detailed illustration can be found in Figure 2. We thus propose our DD loss to effectively measure the discrepancy between

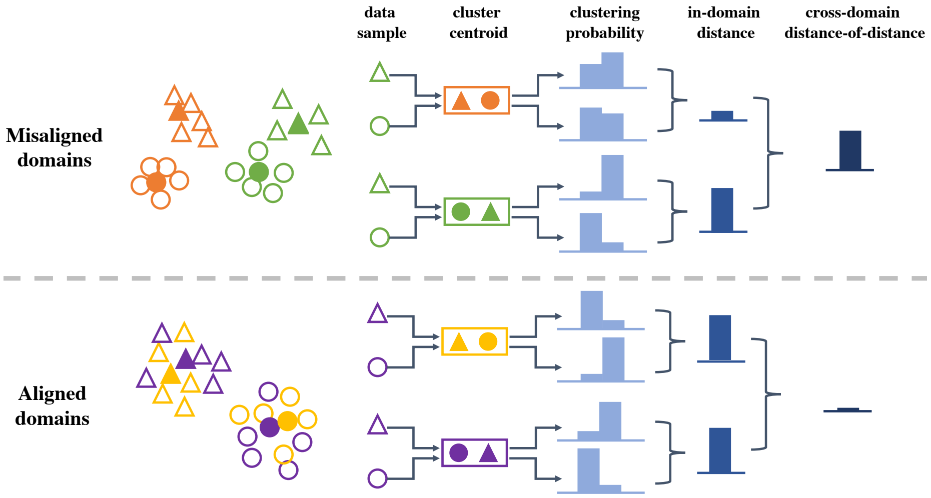

Fig. 2. Illustration of cross-domain distance-of-distance loss. Shapes and colors represent the data samples of different classes and domains, respectively.

features in two domains, where a smaller DD loss indicates better alignment. Formally, our DD loss is written as:

$$\mathcal{L}_{\text{DD}} = \sum_{i \in \mathbf{R}^A} \sum_{j \in \mathbf{R}^A} dd_{ij} + \sum_{i \in \mathbf{R}^B} \sum_{j \in \mathbf{R}^B} dd_{ij}, \tag{8}$$

where $\mathbf{R}^A$ and $\mathbf{R}^B$ contains indices for domain A and domain B data in current batch. i and j are indices for instances.

Entropy minimization. Our DD loss can also be trivially minimized when all the clustering probabilities are uniformly distributed, *i.e.*, all the elements the in $\mathbf{p}_i^A$ and $\mathbf{p}_i^B$ are the same. To prevent this trivial solution, we propose to minimize the self-entropy for all clustering probabilities:

$$\mathcal{L}_{\text{SE}} = \sum_{i \in \mathbf{I}^A} (H(\mathbf{p}_i^A) + H(\mathbf{p}_i^B)) + \sum_{j \in \mathbf{I}^B} (H(\mathbf{p}_j^A) + H(\mathbf{p}_j^B)), \tag{9}$$

where $\mathbf{I}^A$ and $\mathbf{I}^B$ are the indices for all samples in domain A and B, respectively. Consequently, our training objective for cross-domain alignment is given by:

$$\mathcal{L}_{\text{cross-domain}} = \beta \mathcal{L}_{\text{DD}} + \gamma \mathcal{L}_{\text{SE}}, \tag{10}$$

where β and γ are hyper-parameters to balance the two loss terms.

3.4 Summary

To facilitate the feature extractor f_θ training for cross-domain image retrieval without any labeled data as supervision, we introduce a new cluster-wise contrastive learning loss for semantic-aware feature extraction, and propose a novel

DD loss to effectively evaluate whether the two domains are aligned and train the feature extractor f_θ to minimize the DD loss. Our final training target is defined as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{in-domain}} + \mathcal{L}_{\text{cross-domain}}. \tag{11}$$

4 Experiments

4.1 Datasets and Settings

Datasets. Our proposed method is comprehensively evaluated on two datasets: 1) Office-Home [28] offers 4 domains (Art, Clipart, Product, Real) with 65 categories. 2) DomainNet [23] with six different domains (Clipart, Infograph, Painting, Quickdraw, Real and Sketch). We use all six domains and select those categories with more than 200 images in every domain for training and testing. According to this criterion, 7 categories are used in our experiments.

Implementation details. ResNet-50 [11] is employed as the architecture for our feature extractor f_θ. Both features $\mathbf{x}_i$ and cluster centroids $\mathbf{c}_u$ are L2 normalized 128-d vectors. To make sure the whole training procedure is fully unsupervised, we use parameters from the MoCov2 [3] model trained with unlabeled ImageNet dataset [5] to initialize f_θ. The initial learning rate is 0.0002. We follow MoCov2 [3] to use the cosine learning rate schedule that gradually decreases the learning rate to 0. The total number of training epoch is 200. We adopt SGD to update the parameters in the feature extractor with momentum of 0.9 and a batch size of 64. Our framework is built with deep learning library Pytorch [21]. T_1 and T_2 are set to 20 and 100 respectively. The cluster number K is set to be the same as the number of classes in the training set, *i.e.*, 65 for Office-Home and 7 for DomainNet.

Evaluation metrics. To validate the retrieval performance for the Office-Home dataset, we follow [14] to calculate the precision among top 1/5/15 retrieved images as the minimum number of images for one single category is 15. As for DomainNet, we filter out those categories with fewer than 200 images. Thus, we measure the precision for top 50/100/200 retrieved image to provide a more comprehensive evaluation for retrieval accuracy in DomainNet. Since our task is category-level cross-domain retrieval, the retrieved images with the same semantic class as the query are regarded as the correct ones.

Baselines. We use the following works as the baselines to evaluate our proposed method: 1) **ID** [30] achieves unsupervised representation learning by instance discrimination where all instance are well-separated regardless of the category. 2) **ProtoNCE** [15] is a more recent unsupervised representattion learning algorithm which proposes to use prototypes to help encode semantic structure of data and predict more aggregated feature clusters. 3) **CDS** [14] is originally designed for cross-domain self-supervised pre-training. In-domain instance discrimination and cross-domain instance matching are designed for learning a shared embedding space across domains. 4) **PCS** [32] is a cross-domain self-supervised learning approach that uses prototypical contrastive learning for in-domain feature learning and instance-prototype matching for cross-domain alignment.

Table 1. Unsupervised Cross-domain Retrieval Accuracy (%) on Office-Home.

Method	Art→Real			Real→Art			Art→Product		
	P@1	P@5	P@15	P@1	P@5	P@15	P@1	P@5	P@15
ID [30]	35.89	33.13	29.60	39.89	34.42	27.65	25.88	24.91	22.49
ProtoNCE [15]	40.50	36.39	34.00	44.53	39.26	32.99	29.54	27.89	25.75
CDS [14]	45.08	41.15	38.73	44.71	40.75	35.53	32.76	31.47	28.90
PCS [32]	41.70	38.51	36.22	44.96	39.88	33.99	33.29	31.50	29.53
Ours	**45.12**	**42.33**	**40.06**	**47.95**	**43.68**	**38.38**	**35.39**	**34.67**	**32.61**
	Product→Art			Clipart→Real			Real→Clipart		
	P@1	P@5	P@15	P@1	P@5	P@15	P@1	P@5	P@15
ID [30]	32.17	25.94	20.23	29.48	26.48	23.25	35.51	32.17	27.96
ProtoNCE [15]	35.73	30.61	24.55	25.25	22.66	20.83	41.15	37.66	31.95
CDS [14]	35.75	32.48	26.82	32.51	30.30	27.80	38.88	36.48	33.16
PCS [32]	39.24	34.77	28.77	29.07	26.06	24.00	40.60	38.11	34.06
Ours	**42.51**	**37.94**	**31.41**	**33.31**	**30.57**	**28.14**	**44.66**	**41.47**	**37.41**
	Product→Real			Real→Product			Product→Clipart		
	P@1	P@5	P@15	P@1	P@5	P@15	P@1	P@5	P@15
ID [30]	50.73	45.03	39.05	45.12	41.46	38.01	31.52	28.55	24.15
ProtoNCE [15]	53.84	48.25	42.21	47.74	44.85	41.21	36.13	33.99	28.24
CDS [14]	54.00	50.07	45.60	49.39	47.27	43.98	37.69	34.99	30.42
PCS [32]	56.45	50.78	45.37	49.90	47.11	43.73	39.51	**37.51**	32.81
Ours	**57.42**	**52.69**	**47.90**	**51.71**	**48.48**	**44.95**	**42.26**	37.42	**33.74**
	Clipart→Product			Art→Clipart			Clipart→Art		
	P@1	P@5	P@15	P@1	P@5	P@15	P@1	P@5	P@15
ID [30]	24.01	22.42	20.60	26.78	24.79	21.64	21.17	17.86	14.71
ProtoNCE [15]	21.17	20.63	20.47	28.97	26.15	22.98	21.33	17.40	14.46
CDS [14]	27.24	26.46	24.86	25.59	23.77	22.41	22.41	20.34	17.34
PCS [32]	26.39	25.86	24.92	31.23	28.74	26.11	24.51	21.27	17.54
Ours	**27.79**	**27.26**	**25.97**	**32.67**	**30.79**	**28.70**	**27.26**	**23.94**	**20.53**
Average	ID [30]			ProtoNCE [15]			CDS [14]		
	P@1	P@5	P@15	P@1	P@5	P@15	P@1	P@5	P@15
	33.18	29.76	25.78	35.49	32.15	28.30	37.17	34.63	31.30
	PCS [32]			Ours			Improvement		
	P@1	P@5	P@15	P@1	P@5	P@15	P@1	P@5	P@15
	38.07	35.01	31.42	**40.67**	**37.60**	**34.15**	**+2.60**	**+2.59**	**+2.73**

4.2 Results

i) The Office-Home Dataset

Settings. Since there are four domains in the Office-Home dataset, we have altogether 6 different pairs (Art-Real, Art-Product, Clipart-Real, Product-Real, Product-Clipart, Art-Clipart) by matching any two of the four domains. Furthermore, the two domains in one pair can be both regarded as query domain to retrieve images from the other domain.

Results. From the retrieval results in Table 1, we make the following observations: 1) ID [30] and ProtoNCE [15] are designed for single domain feature

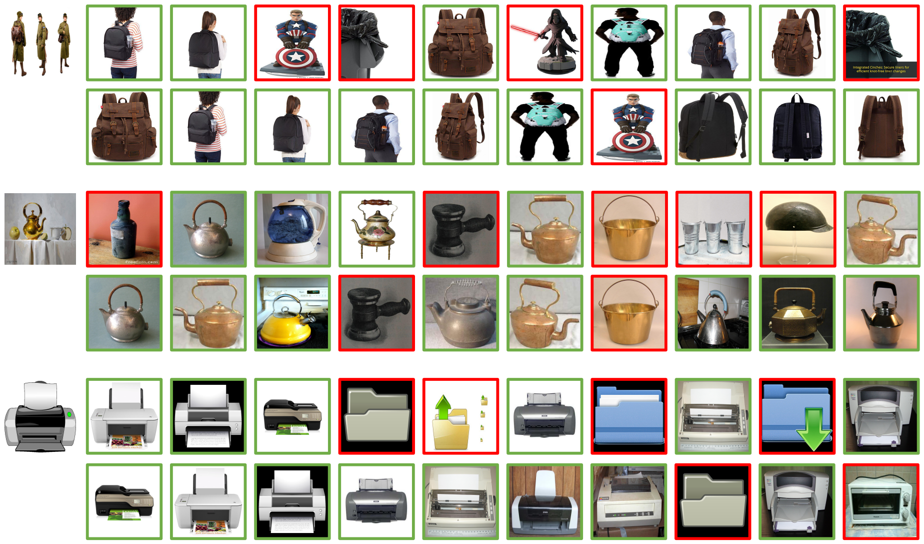

Fig. 3. Top 10 retrieval results on Office-Home. Row 1, 3, 5: Retrieval results of the best baseline method - PCS [32]; Row 2, 4, 6: Retrieval results of our framework. The green and red boxes indicate correct and incorrect retrievals, respectively.

representation learning. The domain gap hurts their performance when applied on cross-domain data. 2) Among all the baseline methods, PCS [32] performs the best. 3) Our proposed method outperforms nearly all the baselines for all pairs in all three evaluation metrics. This shows the effectiveness of our proposed in-domain cluster-wise contrastive learning and the DD loss. 4) The retrieval accuracy is related to the domain gap. The retrieval accuracy is higher when the domain gap of the pair is smaller. In the Office-home dataset, Product and Real are the two with the smallest domain gap. Thus, it can be seen that the retrieval accuracy of both Product → Real and Real → Product are higher than the others. 5) The accuracy for P@1 is always higher than P@5 and P@15, which means it is more likely to retrieve an image from a wrong category when the number of retrieved image becomes larger.

ii) The DomainNet Dataset

Settings. We report the results for 6 pairs of domains from the six domains in DomainNet dataset: Clipart-Sketch, Infograph-Real, Infograph-Sketch, Painting-Clipart, Painting-Quickdraw, Quickdraw-Real. We ensure that every domain appears twice for comprehensive evaluations in all six domains.

Results. The retrieval performance in Table 2 shows that: 1) Similar to the results from Office-Home dataset, PCS [15] is the strongest baseline in our unsupervised cross-domain image retrieval task. 2) Our framework achieves the highest retrieval accuracy for almost all the 12 retrieval tasks. 3) The Quick-

Table 2. Unsupervised Cross-domain Retrieval Accuracy (%) on DomainNet.

Method	Clipart→Sketch			Sketch→Clipart			Infograph→Real		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID [30]	49.46	46.09	40.44	54.38	47.12	37.73	28.27	27.44	26.33
ProtoNCE [15]	46.85	42.67	36.35	54.52	45.04	35.06	28.41	28.53	28.50
CDS [14]	45.84	42.37	37.16	59.13	48.83	37.40	28.51	27.92	27.48
PCS [32]	51.01	46.87	40.19	59.70	50.67	39.38	30.56	30.27	29.68
Ours	**56.31**	**52.74**	**47.38**	**63.07**	**57.26**	**48.17**	**35.52**	**35.24**	**34.35**
	Real→Infograph			Infograph→Sketch			Sketch→Infograph		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID [30]	39.98	31.77	24.84	30.35	29.04	26.55	42.20	34.94	27.52
ProtoNCE [15]	57.01	41.84	30.33	28.24	26.79	24.23	39.83	31.99	24.77
CDS [14]	56.69	39.76	26.38	30.55	**29.51**	**27.00**	**46.27**	36.11	27.33
PCS [32]	55.42	42.13	30.76	30.27	28.36	25.35	42.58	34.09	25.91
Ours	**57.74**	**46.69**	**35.47**	**31.29**	29.33	26.54	43.66	**36.14**	**28.12**
	Painting→Clipart			Clipart→Painting			Painting→Quickdraw		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID [30]	64.67	54.41	40.07	42.37	39.61	35.56	20.34	19.59	18.79
ProtoNCE [15]	55.44	43.74	32.59	39.13	35.87	32.07	21.63	21.24	20.56
CDS [14]	63.15	47.30	32.93	37.75	35.18	32.76	18.75	18.89	17.88
PCS [32]	63.47	53.21	41.68	48.83	46.21	42.10	25.12	24.65	23.80
Ours	**66.42**	**56.84**	**46.72**	**52.58**	**50.10**	**46.11**	**39.72**	**38.59**	**37.63**
	Quickdraw→Painting			Quickdraw→Real			Real→Quickdraw		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID [30]	21.12	19.81	18.48	28.27	27.46	26.32	23.45	22.79	22.01
ProtoNCE [15]	23.95	22.84	21.56	26.38	25.70	24.45	25.10	24.81	23.78
CDS [14]	21.37	21.44	19.46	19.28	19.14	18.67	15.36	15.57	15.82
PCS [32]	24.03	23.24	22.13	34.82	33.92	31.73	28.98	28.85	28.16
Ours	**33.45**	**33.81**	**34.29**	**42.79**	**42.75**	**42.70**	**41.90**	**42.10**	**41.59**
Average	ID [30]			ProtoNCE [15]			CDS [14]		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
	37.07	33.34	28.72	37.21	32.59	27.85	36.89	31.84	26.69
	PCS [32]			Ours			Improvement		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
	41.23	36.87	31.74	**47.09**	**43.47**	**39.09**	**+5.86**	**+6.59**	**+7.35**

draw domain in DomainNet only contain some simple strokes, which lead to the largest domain gap. All the other four baseline methods perform poorly on the Painting → Quickdraw, Quickdraw → Painting, Quickdraw → Real and Real → Quickdraw retrieval. 4) The improvement brought by our proposed method is most significant on Painting → Quickdraw retrieval. Ours is 14.60% higher in terms of P@50 when compared to the best baseline PCS. 5) Among all 12 retrieval pairs, our method performs the best for Painting → Clipart retrieval and achieves 66.42% for P@50.

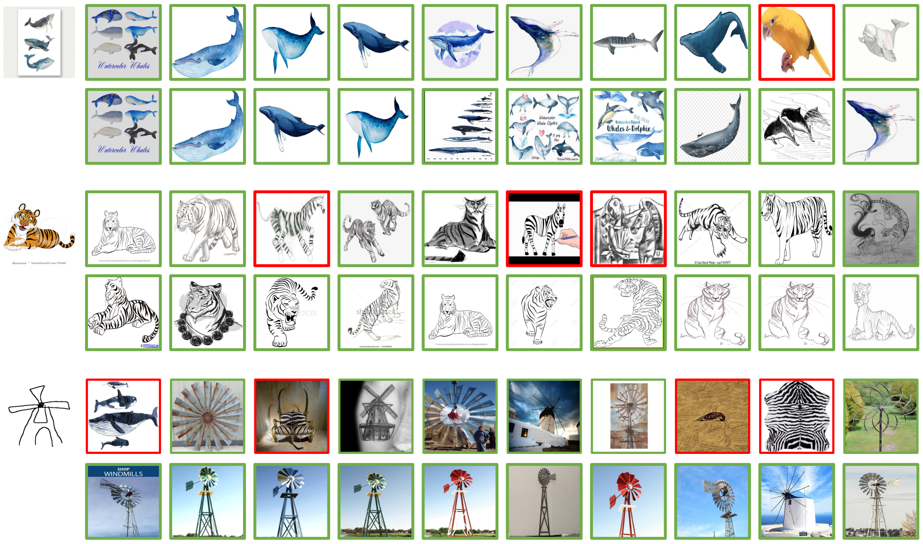

Fig. 4. Top 10 retrieval results on DomainNet. Row 1, 3, 5: Retrieval results of the best baseline method - PCS [32]; Row 2, 4, 6: Retrieval results of our framework. The green and red boxes indicate correct and incorrect retrievals, respectively.

4.3 Ablation Study

The results in Table 3 show the efficacy of different components in our framework. From the table, we can see: 1) Compared to using the instance-wise contrastive learning loss $\mathcal{L}_{IW}$ (v1), our proposed cluster-wise contrastive loss (v2) indeed helps to learn a better feature embedding for cross-domain image retrieval. All three evaluation metrics show the effectiveness of $\mathcal{L}_{CW}$. 2) In v3, we add the self-entropy loss for clustering probabilities. The improvement over v2 shows that the entropy minimization for clustering probabilities is beneficial for cross-domain feature representation learning. 3) Our full model, which employs the DD loss to minimize the discrepancy between domains, provides the best alignment and performs the best compared with all v1, v2 and v3. 4) The efficacy of the DD loss varies a lot in the experiment with different pairs. Comparing to v3 (without DD loss), our full model achieves a performance gain of only 0.66% at P@50 on Real → Infograph, but shows higher performance gain of 8.71% on Real → Quickdraw.

5 Conclusion

This paper presents a novel representation learning framework for unsupervised cross-domain image retrieval which is a challenging but practically valuable task. To extract class semantic-aware feature for category-level retrieval, we propose

Table 3. Ablation Study on our model component. Cross-domain Retrieval Accuracy (%) on DomainNet.

Method	Clipart→Sketch			Sketch→Clipart			Infograph→Real		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
$\mathcal{L}_{IW}$ (v1)	48.08	43.64	37.17	56.09	47.38	37.24	30.51	30.22	29.68
$\mathcal{L}_{IW} + \mathcal{L}_{CW}$ (v2)	51.92	47.95	41.98	60.18	52.01	41.93	32.47	32.00	31.25
$\mathcal{L}_{IW} + \mathcal{L}_{CW} + \mathcal{L}_{SE}$ (v3)	53.17	49.33	43.55	60.66	52.85	42.44	33.80	33.19	32.27
Our full model	**56.31**	**52.74**	**47.38**	**63.07**	**57.26**	**48.17**	**35.52**	**35.24**	**34.35**
	Real→Infograph			Infograph→Sketch			Sketch→Infograph		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
$\mathcal{L}_{IW}$ (v1)	55.85	42.36	30.85	29.90	28.16	25.18	40.84	32.93	25.95
$\mathcal{L}_{IW} + \mathcal{L}_{CW}$ (v2)	57.30	45.51	33.80	30.98	29.21	26.38	**43.81**	35.99	27.86
$\mathcal{L}_{IW} + \mathcal{L}_{CW} + \mathcal{L}_{SE}$ (v3)	57.08	45.88	34.17	30.99	29.25	26.34	43.58	35.83	27.55
Our full model	**57.74**	**46.69**	**35.47**	**31.29**	**29.33**	**26.54**	43.66	**36.14**	**28.12**
	Painting→Clipart			Clipart→Painting			Painting→Quickdraw		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
$\mathcal{L}_{IW}$ (v1)	55.59	45.19	34.46	42.12	38.96	34.50	23.10	22.52	21.47
$\mathcal{L}_{IW} + \mathcal{L}_{CW}$ (v2)	65.00	54.08	41.81	47.66	44.59	40.82	24.09	23.45	22.48
$\mathcal{L}_{IW} + \mathcal{L}_{CW} + \mathcal{L}_{SE}$ (v3)	66.08	**57.20**	**46.88**	**52.71**	49.85	46.05	34.13	33.39	32.24
Our full model	**66.42**	56.84	46.72	52.58	**50.10**	**46.11**	**39.72**	**38.59**	**37.63**
	Quickdraw→Painting			Quickdraw→Real			Real→Quickdraw		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
$\mathcal{L}_{IW}$ (v1)	23.11	22.22	21.20	25.62	24.98	24.05	26.83	26.52	25.53
$\mathcal{L}_{IW} + \mathcal{L}_{CW}$ (v2)	24.83	23.80	22.32	32.32	31.63	30.50	28.25	27.56	26.53
$\mathcal{L}_{IW} + \mathcal{L}_{CW} + \mathcal{L}_{SE}$ (v3)	32.86	32.35	31.45	37.12	37.11	36.63	33.19	33.11	32.54
Our full model	**33.45**	**33.81**	**34.29**	**42.79**	**42.75**	**42.70**	**41.90**	**42.10**	**41.59**

Average											
$\mathcal{L}_{IW}$			$\mathcal{L}_{IW} + \mathcal{L}_{CW}$			$\mathcal{L}_{IW} + \mathcal{L}_{CW} + \mathcal{L}_{SE}$			Our full model		
P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
38.14	33.99	29.42	41.57	37.32	32.31	44.61	40.78	36.01	**47.09**	**43.47**	**39.09**

a cluster-wise contrastive learning loss that pulls samples of similar semantics closer and pushes different clusters apart. For cross-domain alignment, a novel distance of distance loss is introduced to effectively measure the discrepancy between domains and minimized to align features in both domains. The experiment results on Office-Home and DomainNet dataset consistently illustrate the superiority of our proposed algorithm.

Acknowledgements. This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-024), and the Tier 2 grant MOE-T2EP20120-0011 from the Singapore Ministry of Education.

References

1. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
3. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv (2020)
4. Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: ECCV (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
6. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR (2016)
7. Gao, B., Yang, Y., Gouk, H., Hospedales, T.M.: Deep clusteringwith concrete k-means. In: ICASSP (2020)
8. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. ICLR (2018)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014)
10. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. JMLR (2012)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018)
13. Kim, D., Saito, K., Oh, T.H., Plummer, B.A., Sclaroff, S., Saenko, K.: Cross-domain self-supervised learning for domain adaptation with few source labels. arXiv preprint arXiv:2003.08264 (2020)
14. Kim, D., Saito, K., Oh, T.H., Plummer, B.A., Sclaroff, S., Saenko, K.: Cds: Cross-domain self-supervised pre-training. In: ICCV (2021)
15. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. ICLR (2020)
16. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. NeurIPS (2016)
17. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML (2017)
18. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Performance evaluation in content-based image retrieval: overview and proposals. PR (2001)
19. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv (2018)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)
22. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)

23. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
24. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. TOG (2016)
25. Shen, J., Qu, Y., Zhang, W., Yu, Y.: Wasserstein distance guided representation learning for domain adaptation. In: AAAI (2018)
26. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. TPAMI (2000)
27. Song, J., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T.M.: Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: ICCV (2017)
28. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)
29. Villani, C.: Optimal transport: old and new. Springer (2009)
30. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
31. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: CVPR (2016)
32. Yue, X., Zheng, Z., Zhang, S., Gao, Y., Darrell, T., Keutzer, K., Vincentelli, A.S.: Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In: CVPR (2021)
33. Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., Sebe, N.: Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In: CVPR (2021)