Correspondence-Free Domain Alignment for Unsupervised Cross-Domain Image Retrieval

Xu Wang[1], Dezhong Peng[1,3,4], Ming Yan[2], Peng Hu[1*]
[1]College of Computer Science, Sichuan University, Chengdu, China
[2]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3]Sichuan Zhiqian Technology Co., Ltd, Chengdu, China
[4]Chengdu Ruibei Yingte Information Technology Ltd. Company, Chengdu, China
wangxu.scu@gmail.com, pengdz@scu.edu.cn, yanmingtop@gmail.com, penghu.ml@gmail.com

Abstract

Cross-domain image retrieval aims at retrieving images across different domains to excavate cross-domain classificatory or correspondence relationships. This paper studies a less-touched problem of cross-domain image retrieval, *i.e.*, unsupervised cross-domain image retrieval, considering the following practical assumptions: (i) no correspondence relationship, and (ii) no category annotations. It is challenging to align and bridge distinct domains without cross-domain correspondence. To tackle the challenge, we present a novel Correspondence-free Domain Alignment (CoDA) method to effectively eliminate the cross-domain gap through In-domain Self-matching Supervision (ISS) and Cross-domain Classifier Alignment (CCA). To be specific, ISS is presented to encapsulate discriminative information into the latent common space by elaborating a novel self-matching supervision mechanism. To alleviate the cross-domain discrepancy, CCA is proposed to align distinct domain-specific classifiers. Thanks to the ISS and CCA, our method could encode the discrimination into the domain-invariant embedding space for unsupervised cross-domain image retrieval. To verify the effectiveness of the proposed method, extensive experiments are conducted on four benchmark datasets compared with six state-of-the-art methods.

Introduction

With the rapid growth of images collected from many diverse sources (*e.g.*, viewpoints, lightning, artistic styles, and photograph) on the Internet, there are growing demands to develop various applications on different domains, such as domain adaptation (Li et al. 2021b; Singh 2021; Zhu, Zhuang, and Wang 2019), cross-domain clustering (Li et al. 2021a), and cross-domain image retrieval (CIR) (Huang et al. 2015; Wang et al. 2019; Paul, Dutta, and Biswas 2021; Wang et al. 2022; Hu and Lee 2022). In these applications, CIR has attracted more and more attention in recent years for its flexible retrieval ways and achieved great success in numerous application scenarios, *e.g.*, surveillance, mobile product image search (Shen et al. 2012). Given a query image, CIR aims to correctly retrieve relevant images across distinct domains, which are with similar visual information or the same semantics. However, it is challenging to retrieve images across diverse domains due to the inconsistent image distributions, namely the so-called "domain gap" or "cross-domain gap" (Nam et al. 2021).

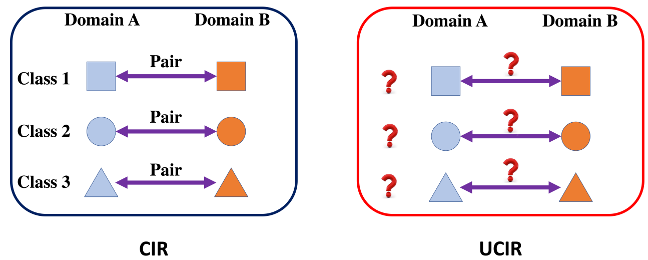


Figure 1: Comparison of cross-domain image retrieval (CIR) and unsupervised cross-domain image retrieval (UCIR). Compared with CIR, there is no category annotation or correspondence relationship in UCIR.

To bridge the domain gap, extensive efforts have been devoted to learning common representations from different domains (Sangkloy et al. 2016; Yu et al. 2016; Sain et al. 2021; Fuentes and Saavedra 2021). Although the existing cross-domain image retrieval methods have achieved promising performance, they implicitly assume that the multi-domain training data are annotated and aligned well. In practice, however, it is extremely expensive and even impossible to label multiple large-scale domains. To alleviate the high labeling cost, one advisable solution is to design an unsupervised cross-domain learning paradigm to learn from a large number of low-cost and highly accessible unlabeled data. Obviously, compared with CIR, unsupervised cross-domain image retrieval (UCIR) is more challenging due to unavailable category and correspondence information as shown in Figure 1. Such a UCIR problem is barely touched so far, to the best of our knowledge.

To overcome the challenges, this paper proposes a novel approach dubbed **Co**rrespondence-free **D**omain **A**lignment (**CoDA**), which unifies the In-domain Self-matching Supervision (ISS) and Cross-domain Classifier Alignment (CCA) to achieve unsupervised cross-domain image retrieval. Specifically, ISS employs a novel self-matching supervision mechanism to encapsulate the discriminative information into the shared embedding space.

[*]Corresponding author.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Meanwhile, different from previous works which mainly focus on the distribution alignment, CCA enforces the predictions by different domain-specific classifiers to be consistent to minimize the cross-domain discrepancy, thus learning the domain-aligned and domain-invariant representations.

The main novelties and contributions of this work are summarized as follows:

- We propose a novel method called **Co**rrespondence-free **D**omain **A**lignment (**CoDA**) to tackle a less-touched problem, *i.e.*, unsupervised cross-domain image retrieval.
- A novel In-domain Self-matching Supervision module (ISS) is proposed to project the discrimination into common representations by simultaneously conducting domain-specific clustering and discriminative learning.
- We present a Cross-domain Classifier Alignment mechanism (CCA) to learn domain-invariant representations by minimizing the discrepancy across domain-specific classifiers.
- Extensive experiments are conducted on four benchmarks, demonstrating the effectiveness of the proposed approach for unsupervised cross-domain image retrieval.

Related Work

Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge learned from a source domain with fully-annotated training examples to a target domain with unlabeled data only. The key challenge of UDA is having to counter the inconsistent distributions across different domains, namely the cross-domain gap. To eliminate the domain gap, discrepancy-based methods are proposed to minimize the Maximum Mean Discrepancy (MMD) (Wynne and Duncan 2022) or Joint MMD (Long et al. 2017) across domains. Motivated by the Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), several GAN-based methods (Tzeng et al. 2017; Hoffman et al. 2018) are also presented to perform domain alignment in feature space. Besides, Saito et al. (2019) propose a novel minimax entropy approach that adversarially achieves domain adaptation. Nevertheless, these UDA methods have supervision on the source domain. Nevertheless, these UDA methods implicitly assume that the source domain is annotated well, which inevitably increases the labeling cost. To tackle this problem, a novel problem is studied in the paper, *i.e.*, the unsupervised cross-domain image retrieval, where both source and target domains are unlabeled.

Cross-Domain Image Retrieval

As an extensively studied task in computer vision, content-based image retrieval (Datta et al. 2008) has been widely explored, where the query and database are constrained to the same domain. In practice, however, we often require retrieving related images across diverse domains, *i.e.*, cross-domain image retrieval (CIR). For example, in online shopping, we need to search for products using images captured by smartphones as the query (Huang et al. 2015), wherein query and database exist in distinct domains. Compared with content-based image retrieval, CIR is more challenging due to the domain gap. To bridge the domain gap, several works exploit the category information (Sangkloy et al. 2016) for discriminative feature extraction or cross-domain pairing by minimizing triplet (Yu et al. 2016) and HOLEF (Song et al. 2017) loss. However, the cross-domain correspondence utilized in these works is labor-intensive, which severely restricts their applications. In this paper, we focus on a more challenging setting, *i.e.*, unsupervised cross-domain image retrieval, which is first introduced in a recent work (Kim et al. 2021) but is still barely touched so far. Different from this work which trains models by contrasting instance-instance pairs, ours employs multi-domain classifiers to capture discrimination and enforce their prediction consistency to alleviate the cross-domain discrepancy, thus embracing better performance.

Methodology

Problem Statement

We first give the formal definition of the unsupervised cross-domain retrieval task. Given two sets of unlabeled training images $\mathcal{D}_A = \{(\mathbf{x}_i^A)\}_{i=1}^{N_A}$ from domain A, and $\mathcal{D}_B = \{(\mathbf{x}_j^B)\}_{j=1}^{N_B}$ from another different domain B, our goal is to learn an effective feature extractor to transform images from different domains into a common embedding space, where the features are discriminative for cross-domain similarity measuring. During the training stage, the training set is utilized to bridge the domain gap. While during testing, a query image $\mathbf{x}_i^A \in \mathcal{D}_A$ with its category label y_i is given, the ultimate goal is to correctly retrieve all semantically similar images in $\mathcal{D}_B$, *i.e.*, all $\mathbf{x}_k^B \in \mathcal{D}_B$ with category label $y_k = y_i$. Under the unsupervised setting, $\mathcal{D}_A$ and $\mathcal{D}_B$ share the same categories, but there are unavailable category annotations or correspondence relationships between $\mathcal{D}_A$ and $\mathcal{D}_B$. This increases the challenge of how to learn a common space from unlabeled and unaligned data for cross-domain retrieval.

Overview

To tackle the above-mentioned challenge, we propose a novel approach named Correspondence-free Domain Alignment (CoDA). We firstly adopt a Convolutional Neural Network (CNN) backbone $f(\cdot; \mathbf{\Theta})$ pretrained on ImageNet to extract the features. Formally, given an image point $\mathbf{x}_i$, the embedded feature $\mathbf{v}_i$ could be obtained by:

$$\mathbf{v}_i = f(\mathbf{x}_i; \mathbf{\Theta}) \in \mathbb{R}^L, \tag{1}$$

where $\mathbf{\Theta}$ is the parameter set of CNN. L denotes the dimensionality of the common space.

As illustrated in Figure 2, our CoDA contains two parts:

(1) In-domain Self-matching Supervision (ISS). By performing domain-specific clustering and in-domain self-matching with the soft label, ISS could encapsulate the discriminative information of image data into the embedding space.

(2) Cross-domain Classifier Alignment (CCA). By minimizing the discrepancy among different domain-specific

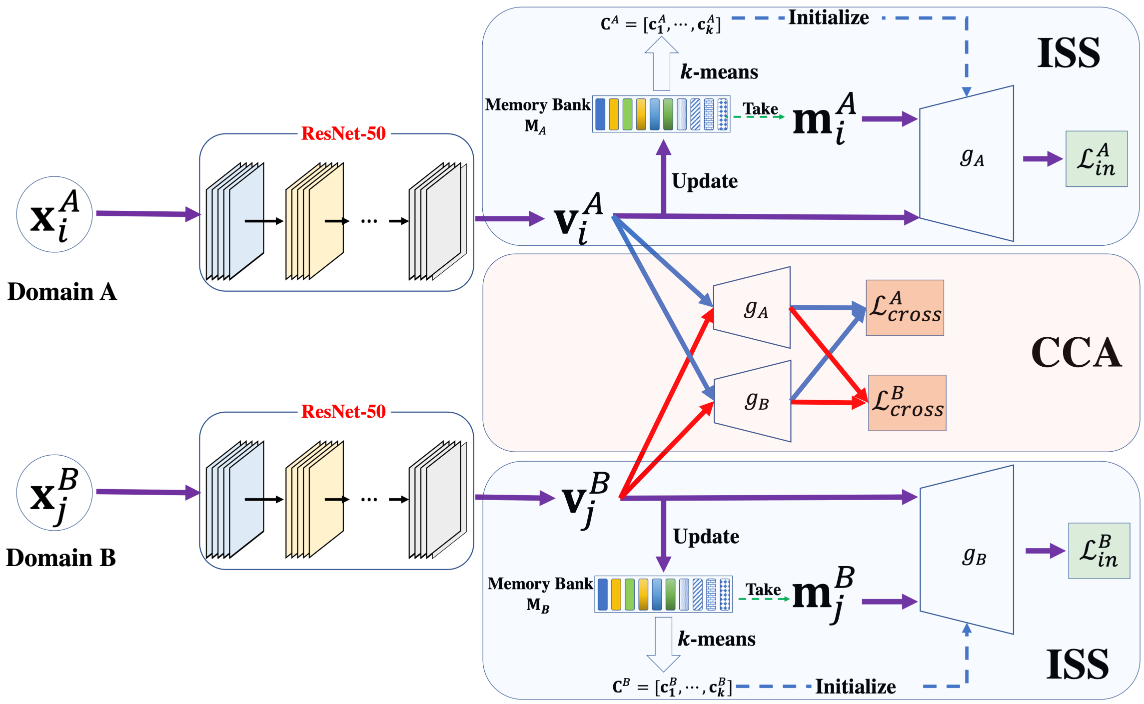


Figure 2: The pipeline of our CoDA for unsupervised cross-domain image retrieval. The ResNet-50 network is adopted to learn the feature embeddings for different domains. ISS performs domain-specific clustering and in-domain self-matching with the soft-label, thus encapsulating the discrimination into the embedding space. CCA minimizes cross-domain discrepancy by enforcing the predictions of different domain-specific classifiers to be consistent.

classifiers, CCA achieves domain alignment through enforcing the predictions by different domain-specific classifiers to be consistent.

To learn discriminative common representations, our CoDA unifies ISS and CCA to encapsulate discrimination while eliminating the cross-domain discrepancy in the latent common space. The overall objective function could be formulated as:

$$\mathcal{L}_{join} = \mathcal{L}_{in} + \lambda \mathcal{L}_{cross}, \quad (2)$$

where $\mathcal{L}_{in}$ and $\mathcal{L}_{cross}$ are the objectives of ISS and CCA, respectively. λ is the parameter to balance the contribution of $\mathcal{L}_{cross}$. Due to the randomness of clustering, we perform k-means on the samples R times with different numbers of clusters $\{k_r\}_{r=1}^{R}$ for relatively stable results. In summary, the objective function of CoDA can be rewritten as:

$$\mathcal{L} = \frac{1}{R} \sum_{r=1}^{R} \mathcal{L}_{join}^{(r)} \quad (3)$$

To train the proposed method, we adopt a gradient descent optimizer to minimize the objective function in a batch-by-batch manner. Algorithm 1 briefly summarizes the optimization procedure of the proposed CoDA approach.

In-Domain Self-Matching Supervision

In UCIR task, it is important to learn discriminative features. Intuitively, we expect that the features within the same cluster stay close while those in different clusters are apart from each other. To this end, we propose a novel In-domain Self-matching Supervision (ISS) module. There are two key steps in our ISS: (1) domain-specific clustering and (2) in-domain self-matching with the soft label.

(1) Domain-specific Clustering. We firstly maintain two memory banks $\mathbf{M}_A$ and $\mathbf{M}_B$ for $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively:

$$\mathbf{M}_A = [\mathbf{m}_1^A, \cdots, \mathbf{m}_{N_A}^A], \mathbf{M}_B = [\mathbf{m}_1^B, \cdots, \mathbf{m}_{N_B}^B]. \quad (4)$$

The memory banks are initialized with the features extracted by $f(\cdot; \mathbf{\Theta})$. During training, the features in memory banks $\mathbf{M}_A$ and $\mathbf{M}_B$ are updated with a momentum η after every batch:

$$\begin{aligned} \mathbf{m}_i^A &= \eta \mathbf{m}_i^A + (1-\eta)\mathbf{v}_i^A, \\ \mathbf{m}_j^B &= \eta \mathbf{m}_j^B + (1-\eta)\mathbf{v}_j^B. \end{aligned} \quad (5)$$

To reduce the high memory cost of gradient computation, the memory banks are updated without producing gradients. After initialization, we then respectively perform k-means clustering on $\mathbf{M}_A$ and $\mathbf{M}_B$ to obtain domain-specific centroids $\mathbf{C}^A = [\mathbf{c}_1^A, \cdots, \mathbf{c}_k^A]$ and $\mathbf{C}^B = [\mathbf{c}_1^B, \cdots, \mathbf{c}_k^B]$. Note that, we use the union of the features from two domains and perform k-means to get global clustering centroids, which are leveraged as the initialized centroids of domain-specific clustering. Meanwhile, different from Deep Clustering (Caron et al. 2018) which conducts k-means every epoch, our method only conducts k-means clustering once in the first epoch, thus embracing higher efficiency.

(2) In-domain Self-matching with Soft Label. In order to learn discriminative representations for each domain, we design two classifiers $\sigma(g_A(\cdot))$ and $\sigma(g_B(\cdot))$ for $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively, where $\sigma(\cdot)$ denotes the softmax function, $g_A(\cdot)$ and $g_B(\cdot)$ are two linear layers with weights $\mathbf{W}_A$ and $\mathbf{W}_B$, which are respectively initialized with domain-specific centroids $\mathbf{C}^A$ and $\mathbf{C}^B$. Since there are unavailable labels in unsupervised cross-domain image retrieval, we need a self-labeling mechanism to assign the labels automatically. During the self-labeling, the memory bank $\mathbf{M}_A$ is initialized with $\mathbf{v}_i^A$. Based on the consistency regularization (Bachman, Alsharif, and Precup 2014; Sohn et al. 2020), which holds that the model should output similar predictions when fed augmented versions of the same image, we regard the memory feature $\mathbf{m}_i^A$ as an augmentation of $\mathbf{v}_i^A$, proposing to achieve in-domain self-matching via the loss function:

$$\mathcal{L}_{in}^A = \frac{1}{N_A}\sum_{i=1}^{N_A} \mathrm{H}\left(\sigma\left(\frac{g_A(\mathbf{m}_i^A)}{\tau}\right), \sigma\left(g_A(\mathbf{v}_i^A)\right)\right), \quad (6)$$

where $\sigma(\frac{g_A(\mathbf{m}_i^A)}{\tau})$ is the soft label of image $\mathbf{x}_i^A$, and $\mathrm{H}(p,q)$ denotes the cross-entropy between two probability distributions p and q, and τ is the temperature parameter. Similarly, we have the loss for domain $\mathcal{D}_B$ as $\mathcal{L}_{in}^B$:

$$\mathcal{L}_{in}^B = \frac{1}{N_B}\sum_{j=1}^{N_B} \mathrm{H}\left(\sigma\left(\frac{g_B(\mathbf{m}_j^B)}{\tau}\right), \sigma\left(g_B(\mathbf{v}_j^B)\right)\right). \quad (7)$$

Then the loss of ISS can be written as:

$$\mathcal{L}_{in} = \mathcal{L}_{in}^A + \mathcal{L}_{in}^B. \quad (8)$$

Cross-Domain Classifier Alignment

With ISS, the model is supposed to learn discriminative features for each domain. However, it ignores the domain invariance, which is another important demand for cross-domain image retrieval. In an unsupervised setting, it is challenging to learn domain-invariant features since there is no correspondence between different domains. To encourage domain-aligned as well as domain-invariant features across different domains, we propose a Cross-domain Classifier Alignment mechanism (CCA), which minimizes the discrepancy between different domain-specific classifiers.

For our proposal, we hold that the classifiers which are trained on different domains have a disagreement on the predictions of the same feature. The disagreement is especially obvious when the features are near the class boundaries. To learn domain-invariant features across different domains, the predictions by domain-specific classifiers of the same image should be consistent. Thus, the discrepancy across different domain-specific classifiers is supposed to be minimized.

Concretely, take the feature $\mathbf{v}_i^A$ as example, which is feed into the domain-specific classifiers $g_A(\cdot)$ and $g_B(\cdot)$ to get logits. Then we employ the mean absolute values of the difference between the logits of different domain-specific classifiers as the cross-domain alignment loss:

$$\mathcal{L}_{cross}^A = \frac{1}{N_A}\sum_{i=1}^{N_A} \left|g_A(\mathbf{v}_i^A) - g_B(\mathbf{v}_i^A)\right|. \quad (9)$$

Algorithm 1: Optimization procedure of CoDA

Input: The training dataset $\mathcal{D}_A = \{(\mathbf{x}_i^A)\}_{i=1}^{N_A}$ from domain A, and $\mathcal{D}_B = \{(\mathbf{x}_j^B)\}_{j=1}^{N_B}$ from domain B, the dimensionality of the common space L, memory bank update momentum η, batch size n_b, balance parameter λ, maximal epoch number N_e, numbers of clustering $\{k_r\}_{r=1}^R$, temperature parameter τ, and initial learning rate α.

Calculate the features for all images from both domains by using the backbone network $f(\cdot, \mathbf{\Theta})$ according to Equation (1).
Initialize memory banks $\mathbf{M}_A$ and $\mathbf{M}_B$ with the calculated features.
Perform domain-specific clustering and initialize the domain-specific classifiers with domain-specific centroids.
for $1, 2, \cdots, N_e$ **do**
  repeat
    Randomly select n_b images from $\mathcal{D}_A$ and n_b images from $\mathcal{D}_B$ to construct mini-batch data.
    Calculate the representations for all images of the mini-batch by using the backbone network $f(\cdot, \mathbf{\Theta})$ according to Equation (1).
    Compute $\mathcal{L}_{in}$ and $\mathcal{L}_{cross}$ according to Equations (8) and (11) on the mini-batch, respectively.
    Update network parameters $\mathbf{\Psi} = \{\mathbf{\Theta}, \mathbf{W}_A, \mathbf{W}_B\}$ by minimizing $\mathcal{L}$ in Equation (3) with descending their stochastic gradient.
    Update memory banks according to Equation (5).
  until all images are selected
end for

Output: Optimized network parameters $\{\mathbf{\Theta}, \mathbf{W}_A, \mathbf{W}_B\}$.

Similarly, we have the cross-domain alignment loss for $\mathbf{v}_j^B$ as follows:

$$\mathcal{L}_{cross}^B = \frac{1}{N_B}\sum_{j=1}^{N_B} \left|g_A(\mathbf{v}_j^B) - g_B(\mathbf{v}_j^B)\right|. \quad (10)$$

Finally, the loss of CCA can be written as:

$$\mathcal{L}_{cross} = \mathcal{L}_{cross}^A + \mathcal{L}_{cross}^B. \quad (11)$$

Experiments

Datasets

To verify the effectiveness of the proposed method, we conduct extensive experiments on four benchmark datasets, *i.e.*, Office31 (Saenko et al. 2010), Image-CLEF (Long et al. 2017), OfficeHome (Venkateswara et al. 2017), and Adaptiope (Ringwald and Stiefelhagen 2021). For each dataset, we randomly partition the data into training and test sets, with an 80-20 ratio for each category. **Office31**: This dataset consists of three real-world object domains: Amazon (A), Webcam (W), and DSLR (D). It has 4,652 images with 31 categories. We conduct six retrieval tasks: A-D, A-W, D-A, D-W, W-A, W-D. **Image-CLEF**: The dataset is a benchmark dataset for ImageCLEF 2014 domain adaptation challenge.

Method		Cross-domain Retrieval Task on Office31 and Adaptiope datasets													
		Office31							Adaptiope						
		A-D	A-W	D-A	D-W	W-A	W-D	Avg	P-R	P-S	R-P	R-S	S-P	S-R	Avg
ImageNet	/	0.569	0.500	0.617	0.816	0.552	0.801	0.643	0.400	0.191	0.395	0.137	0.222	0.165	0.252
MMD (Wynne and Duncan 2022)	Best	0.426	0.356	0.524	0.716	0.450	0.709	0.529	0.274	0.100	0.244	0.066	0.093	0.090	0.145
	Last	0.146	0.126	0.460	0.673	0.319	0.636	0.393	0.016	0.014	0.013	0.013	0.017	0.014	0.015
SimCLR (Chen et al. 2020)	Best	0.544	0.496	0.617	0.819	0.543	0.812	0.638	0.362	0.202	0.343	0.135	0.211	0.156	0.235
	Last	0.540	0.473	0.614	0.807	0.535	0.804	0.629	0.254	0.135	0.248	0.091	0.160	0.117	0.168
InstDis (Wu et al. 2018)	Best	0.509	0.452	0.640	0.847	0.570	0.812	0.638	0.429	0.222	0.418	0.164	0.241	0.183	0.276
	Last	0.329	0.263	0.520	0.704	0.434	0.602	0.475	0.331	0.151	0.325	0.114	0.170	0.135	0.204
CDS(Kim et al. 2021)	Best	0.667	0.625	0.709	0.900	0.644	0.884	0.738	0.575	0.352	0.574	0.250	0.361	0.254	0.394
	Last	0.520	0.478	0.693	0.864	0.624	0.796	0.663	0.540	0.328	0.549	0.212	0.333	0.228	0.365
PCS (Yue et al. 2021)	Best	**0.727**	0.707	**0.753**	0.885	0.712	0.892	0.779	0.569	0.348	**0.583**	0.270	0.337	0.259	0.394
	Last	**0.711**	0.692	0.742	0.875	0.706	0.886	0.769	0.556	0.340	0.561	**0.256**	0.332	0.244	0.382
CoDA (ours)	Best	0.717	**0.714**	0.749	**0.914**	**0.731**	**0.902**	**0.788**	**0.598**	**0.376**	0.582	**0.286**	**0.393**	**0.301**	**0.423**
	Last	0.709	**0.698**	**0.743**	**0.914**	**0.721**	**0.901**	**0.781**	**0.587**	**0.343**	**0.574**	0.254	**0.383**	**0.281**	**0.404**

Table 1: The mAP@All retrieval performance comparison for the proposed CoDA approach and other compared methods on Office31 dataset. The best and second best performance results among all methods are in bold and in underline, *resp*.

Method		Cross-domain Retrieval Task on ImageCLEF dataset												
		B-C	B-I	B-P	C-B	C-I	C-P	I-B	I-C	I-P	P-B	P-C	P-I	Avg
ImageNet	/	0.553	0.498	0.432	0.538	0.754	0.654	0.479	0.742	0.635	0.395	0.640	0.619	0.578
MMD (Wynne and Duncan 2022)	Best	0.496	0.466	0.363	0.504	0.714	0.558	0.487	0.705	0.554	0.371	0.540	0.558	0.526
	Last	0.442	0.414	0.312	0.453	0.620	0.476	0.430	0.600	0.474	0.323	0.460	0.487	0.458
SimCLR (Chen et al. 2020)	Best	0.552	0.508	0.431	0.537	0.790	0.667	0.498	0.793	0.647	0.410	0.670	0.640	0.595
	Last	0.543	0.503	0.426	0.534	0.780	0.661	0.485	0.792	0.640	0.410	0.669	0.631	0.590
InstDis (Wu et al. 2018)	Best	0.521	0.492	0.418	0.530	0.748	0.638	0.487	0.718	0.641	0.404	0.600	0.608	0.567
	Last	0.376	0.373	0.314	0.365	0.477	0.383	0.376	0.462	0.470	0.318	0.373	0.455	0.395
CDS (Kim et al. 2021)	Best	0.643	0.632	0.532	0.643	0.912	0.778	0.627	**0.910**	0.796	0.526	0.747	0.768	0.709
	Last	0.529	0.541	0.440	0.529	0.779	0.664	0.509	0.787	0.656	0.394	0.652	0.609	0.591
PCS (Yue et al. 2021)	Best	0.673	0.635	0.534	0.688	**0.918**	0.777	0.619	0.903	0.776	0.535	0.749	0.770	0.714
	Last	0.625	0.609	0.496	0.661	**0.900**	0.767	0.576	0.886	0.745	0.461	0.745	0.770	0.687
CoDA (ours)	Best	**0.688**	**0.656**	**0.556**	**0.705**	0.901	**0.785**	**0.661**	**0.910**	**0.784**	**0.552**	**0.752**	**0.773**	**0.727**
	Last	**0.685**	**0.656**	**0.556**	**0.700**	0.880	**0.774**	**0.660**	**0.900**	**0.778**	**0.550**	**0.749**	**0.771**	**0.721**

Table 2: The mAP@All retrieval performance comparison for the proposed CoDA approach and other compared methods on ImageCLEF dataset. The best and second best performance results among all methods are in bold and in underline, *resp*.

It is composed by selecting the 12 common classes shared by four public domains: Bing (B), Caltech256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). For each domain, there are 50 images in each category. We conduct 12 retrieval tasks on this dataset. **OfficeHome**: This dataset contains four domains where each domain consists of 65 categories. The four domains are Artistic images (A), Clipart (C), Product images (P), and Real-world images (R). It contains 15,500 images, with an average of around 70 images per class and a maximum of 99 images in a category. We also conduct 12 retrieval tasks for this dataset. **Adaptiope**: Adaptiope is a dataset that offers 123 classes in three different domains. The domains are Synthetic (S), Product (P), and Real life (R). There are totally 36,900 images, with a maximum of 100 images in a category. We also conduct six retrieval tasks for this dataset.

Implementation Detail

In CoDA, a ResNet-50 network pre-trained on ImageNet is utilized to initialize the backbone. Meanwhile, we replace the last FC layer with a 512-D randomly initialized linear layer. The features are ℓ_2-normalized. To obtain stable performance, four k-means are conducted on the obtained features, of which cluster numbers respectively are $\{n_k, 2n_k, 3n_k, 4n_k\}$, where n_k could be empirically set as 50 (for Office31 and ImageCLEF datasets) and 100 (for OfficeHome and Adaptiope datasets). Stochastic Gradient Descent (SGD) optimizer is adopted to train our CoDA. For a fair comparison, the hyper-parameters are set as $\eta = 0.95$ $n_b = 16$, $\lambda = 0.01$, $N_e = 20$, $\tau = 0.01$, and $\alpha = 0.003$ for all datasets. The proposed approach is implemented by PyTorch with two Nvidia GeForce RTX 2080 GPUs.

Experimental Setup

In the experiments, we evaluate the effectiveness of the proposed approach compared with several state-of-the-art baselines. The compared methods are as follows: 1) **ImageNet** is a commonly-used baseline that is trained on ImageNet. 2) **MMD** (Wynne and Duncan 2022) is a kernel-based approach aimed at measuring the distance between two probability distributions in a reproducing kernel Hilbert space. 3) **InstDis** (Wu et al. 2018) exploits instance discrimination to achieve unsupervised representation learning. 4) **SimCLR** (Chen et al. 2020) learns representations by maximizing agreement between differently augmented views of the same example via a contrastive loss in the latent space. 5) **CDS** (Kim et al. 2021) is designed for cross-domain self-supervised pre-training. 6) **PCS** (Yue et al. 2021) is a cross-domain self-supervised learning method for few-shot unsupervised domain adaptation. We adopt mean average pre-

Method		Cross-domain Retrieval Task on OfficeHome dataset												
		A-C	A-P	A-R	C-A	C-P	C-R	P-A	P-C	P-R	R-A	R-C	R-P	Avg
ImageNet	/	0.200	0.325	0.387	0.196	0.246	0.268	0.329	0.261	0.478	0.362	0.247	0.438	0.311
MMD (Wynne and Duncan 2022)	Best	0.163	0.252	0.325	0.150	0.170	0.208	0.243	0.187	0.358	0.295	0.208	0.354	0.243
	Last	0.083	0.112	0.155	0.075	0.039	0.054	0.137	0.046	0.087	0.148	0.062	0.080	0.090
SimCLR (Chen et al. 2020)	Best	0.211	0.314	0.374	0.191	0.229	0.265	0.299	0.251	0.459	0.347	0.257	0.440	0.303
	Last	0.200	0.300	0.364	0.165	0.196	0.226	0.282	0.212	0.429	0.329	0.222	0.424	0.279
InstDis (Wu et al. 2018)	Best	0.242	0.348	0.395	0.220	0.253	0.277	0.334	0.277	0.477	0.365	0.271	0.447	0.326
	Last	0.221	0.278	0.331	0.173	0.191	0.216	0.203	0.205	0.326	0.268	0.218	0.335	0.247
CDS (Kim et al. 2021)	Best	0.330	0.44.5	0.514	0.324	0.403	0.418	0.452	0.415	0.608	0.511	0.420	0.588	0.452
	Last	0.327	0.438	0.492	0.282	0.360	0.381	0.402	0.377	0.542	0.441	0.397	0.523	0.414
PCS (Yue et al. 2021)	Best	0.343	0.463	0.516	0.323	0.405	0.406	0.470	0.421	0.613	0.516	0.428	0.601	0.459
	Last	0.335	0.458	0.513	0.306	0.390	0.398	0.452	0.412	0.605	0.492	0.405	0.597	0.447
CoDA (ours)	Best	**0.347**	**0.496**	**0.532**	**0.332**	**0.429**	**0.447**	**0.504**	**0.452**	**0.652**	**0.531**	**0.460**	**0.652**	**0.486**
	Last	**0.347**	**0.494**	**0.530**	**0.329**	**0.421**	**0.440**	**0.502**	**0.446**	**0.648**	**0.531**	**0.457**	**0.650**	**0.482**

Table 3: The mAP@All retrieval performance comparison for the proposed CoDA approach and other compared methods on OfficeHome dataset. The best and second best performance results among all methods are in bold and in underline, *resp*.

Method		Cross-domain Retrieval Task on OfficeHome dataset												
		A-C	A-P	A-R	C-A	C-P	C-R	P-A	P-C	P-R	R-A	R-C	R-P	Avg
CoDA (with $\mathcal{L}_{in}$ only)	Best	0.323	0.464	0.518	0.313	0.377	0.411	0.476	0.403	0.637	0.537	0.422	0.626	0.459
	Last	0.320	0.463	0.518	0.311	0.374	0.395	0.476	0.390	0.628	0.537	0.408	0.613	0.453
CoDA (with $\mathcal{L}_{cross}$ only)	Best	0.114	0.182	0.238	0.100	0.074	0.097	0.179	0.092	0.196	0.225	0.099	0.149	0.145
	Last	0.101	0.141	0.158	0.084	0.041	0.054	0.145	0.047	0.149	0.158	0.036	0.075	0.099
CoDA (full)	Best	**0.347**	**0.496**	**0.532**	**0.332**	**0.429**	**0.447**	**0.504**	**0.452**	**0.652**	**0.531**	**0.460**	**0.652**	**0.486**
	Last	**0.347**	**0.494**	**0.530**	**0.329**	**0.421**	**0.440**	**0.502**	**0.446**	**0.648**	**0.531**	**0.457**	**0.650**	**0.482**

Table 4: The mAP@All retrieval performance comparison for the CoDA (full version) and its two variants on OfficeHome dataset. The best performance results among all methods are in bold.

cision on all retrieved results (mAP@All) as the evaluation metric to measure the performance of the methods. For a fair and comprehensive comparison, we report the best and last mAP@All results among all epochs.

Comparison With State-of-the-Art Methods

We conduct unsupervised cross-domain image retrieval on the four datasets to evaluate the performance of our CoDA and the compared methods. The experimental results are reported in Tables 1, 2, and 3. From the experimental results, one could obtain the following observations: (1) Our CoDA outperforms other methods on all datasets, in almost all retrieval cases. The results demonstrate the superiority of the proposed method for unsupervised cross-domain image retrieval. For example, in Table 2 and Table 3, our CoDA respectively surpasses PCS by 1.8% and 5.9% in terms of average mAP@All. The reason is that our CoDA can eliminate the cross-domain gap by encapsulating the discrimination into the domain-invariant embedding space. (2) Self-supervised representation learning methods are specifically designed for the single-domain task, and cannot achieve satisfactory performance for cross-domain retrieval, *e.g.*, InstDis (Wu et al. 2018) and SimCLR (Chen et al. 2020). The results indicate that the cross-domain gap impedes their performance when applied to multi-domain data. (3) Compared to self-supervised representation learning methods, CDS (Kim et al. 2021) and PCS (Yue et al. 2021) achieves better performance. The results indicate that exploiting both in- and cross-domain learning could boost the performance of unsupervised cross-domain image retrieval.

Ablation Study

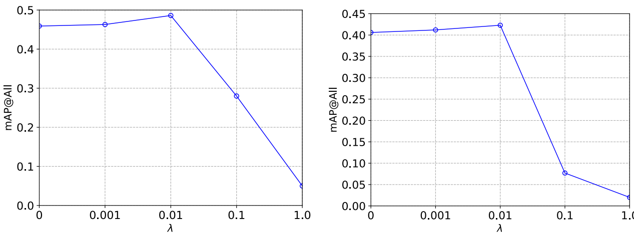
In this section, we evaluate the contributions of the proposed components (*i.e.*, $\mathcal{L}_{in}$ and $\mathcal{L}_{cross}$) for unsupervised cross-domain image retrieval. To this end, we compare our CoDA with its two variations (*i.e.*, CoDA with $\mathcal{L}_{in}$ only and CoDA with $\mathcal{L}_{cross}$ only) on the OfficeHome dataset. The experimental results are shown in Table 4. From the table, one could observe that the performance of our CoDA will degrade significantly in the absence of $\mathcal{L}_{in}$ or $\mathcal{L}_{cross}$, which demonstrates that both the two losses contribute to unsupervised cross-domain image retrieval in our framework. Furthermore, the method will fail to work without L_{in}, demonstrating that in-domain discrimination excavation is crucial for cross-domain retrieval.

Effect of Coefficient λ

To investigate the impact of the coefficient λ in Eq. (3), we conduct parameter analysis experiments on OfficeHome and Adaptiope as shown in Figure 3. The figure plots mAP@all scores *w.r.t.* different values of λ. In the figure, one could find that setting $\lambda = 0.01$ achieves the best performance on both OfficeHome and Adaptiope. Based on the observation, we set $\lambda = 0.01$ in all experiments.

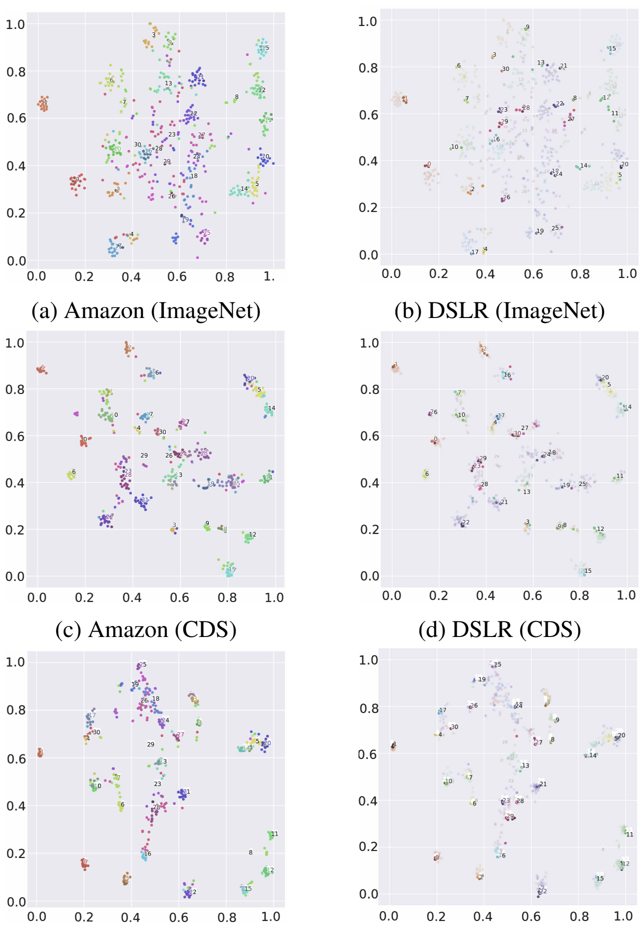
Visualization of the Learned Embeddings

To demonstrate the discriminative information of the learned embeddings, we plot the learned embeddings of ImageNet-pretraining, CDS, and CoDA with t-SNE (Van der Maaten and Hinton 2008) on the Amazon-to-DSLR setting in Office31. The illustrations are shown in Figure 4. From the fig-



(a) OfficeHome (b) Adaptiope

Figure 3: Unsupervised cross-domain image retrieval performance of CoDA in terms of mAP@All scores versus different values of λ on OfficeHome and Adaptiope datasets.



(a) Amazon (ImageNet) (b) DSLR (ImageNet)

(c) Amazon (CDS) (d) DSLR (CDS)

(e) Amazon (CoDA) (f) DSLR (CoDA)

Figure 4: 2-d t-SNE visualizations using 512-d feature representations learned by ImageNet pre-pretraining, CDS, and the proposed CoDA on the testing set of Office31. Each sample is represented by a marker and colored by its corresponding label. *Best viewed in color.*

ure, one could observe that: (1) Compared with ImageNet-pretraining and CDS, the proposed approach well clusters the features with the same class from both domains, demonstrating that our CoDA favors more discriminative features.

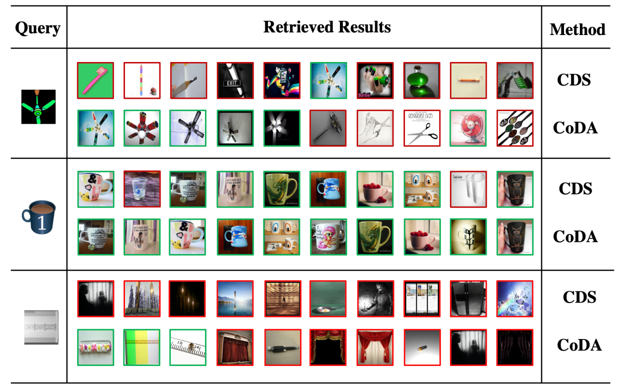


Figure 5: Top-10 UCIR results obtained by CDS and CoDA on OfficeHome dataset. Retrieval is performed by nearest neighbors search using cosine distance on 512-d real value feature vectors. Green borders denote correctly retrieved results, and the red borders demonstrate incorrect retrieved candidates.

(2) The features of CoDA from the two domains are well aggregated, which demonstrates that CoDA learns better domain-alignment for different domains.

Example of Retrievals

In Figure 5, we visually show the top 10 unsupervised cross-domain image retrieval results using CDS and CoDA approaches. The retrievals are conducted on the OfficeHome with 512-D real value features for task C-A (Clipart-Art). The red borders indicate wrongly retrieved results while the green borders denote correctly retrieved results. From these examples, we can observe that the proposed method obtains promising results in most cases compared to CDS. For example, in the third query, given the "ruler" query, CDS fails to retrieve correct images, while our CoDA could find the correct ones in top-3 retrievals. In the wrongly retrieved results, CoDA fails to search for the correct images since the queries and the retrieved images are visually similar. For instance, given the "fan" image (in the first query), a "scissor" image is wrongly retrieved, which shares some visual similarities with the "fan" image.

Conclusion

This paper proposes a novel method called CoDA for unsupervised cross-domain image retrieval. The proposed method is designed to bridge the domain gap through In-domain Self-matching Supervision (ISS) and Cross-domain Classifier Alignment (CCA). Specifically, ISS encapsulates discriminative information into common representations by self-matching. Meanwhile, CCA aligns the features from different domains by minimizing the discrepancy among different domain-specific classifiers. Thanks to ISS and CCA, CoDA enjoys more discriminability in the common space and embraces the domain invariance in unsupervised cross-domain image retrieval. Experimental results on four benchmarks have verified the effectiveness of CoDA.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. U19A2078, 61971296, 62102274), China Postdoctoral Science Foundation (Nos. 2021TQ0223, 2022M712236, 2021M692270), Sichuan Science and Technology Planning Project (Grant Nos. 2023ZHCG0016, 2023YFG0033, 2022YFQ0014, 2022YFH0021, 2021YFS0389, 2021YFS0390), and Fundamental Research Funds for the Central Universities (No. 2022SCU12081).

References

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with Pseudo-Ensembles. In *Advances in Neural Information Processing Systems*, volume 27.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of the European Conference on Computer Vision*, 132–149.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2): 1–60.

Fuentes, A.; and Saavedra, J. M. 2021. Sketch-qnet: A Quadruplet Convnet for Color Sketch-based Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2134–2141.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent Adversarial Domain Adaptation. In *International Conference on Machine Learning*, 1989–1998. Pmlr.

Hu, C.; and Lee, G. H. 2022. Feature Representation Learning for Unsupervised Cross-domain Image Retrieval. In *Proceedings of the European Conference on Computer Vision*.

Huang, J.; Feris, R. S.; Chen, Q.; and Yan, S. 2015. Cross-domain Image Retrieval with A Dual Attribute-aware Ranking Network. In *Proceedings of the IEEE International Conference on Computer Vision*, 1062–1070.

Kim, D.; Saito, K.; Oh, T.-H.; Plummer, B. A.; Sclaroff, S.; and Saenko, K. 2021. CDS: Cross-domain Self-supervised Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9123–9132.

Li, J.; Li, G.; Shi, Y.; and Yu, Y. 2021a. Cross-domain Adaptive Clustering for Semi-supervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2505–2514.

Li, S.; Lv, F.; Xie, B.; Liu, C. H.; Liang, J.; and Qin, C. 2021b. Bi-Classifier Determinacy Maximization for Unsupervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8455–8464.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *International Conference on Machine Learning*, 2208–2217. PMLR.

Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2021. Reducing Domain Gap by Reducing Style Bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8690–8699.

Paul, S.; Dutta, T.; and Biswas, S. 2021. Universal Cross-domain Retrieval: Generalizing Across Classes and Domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12056–12064.

Ringwald, T.; and Stiefelhagen, R. 2021. Adaptiope: A Modern Benchmark for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 101–110.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting Visual Category Models to New Domains. In *Proceedings of the European Conference on Computer Vision*, 213–226. Springer.

Sain, A.; Bhunia, A. K.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2021. Stylemeup: Towards Style-agnostic Sketch-based Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8504–8513.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019. Semi-supervised Domain Adaptation via Minimax Entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8050–8058.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Transactions on Graphics (TOG)*, 35(4): 1–12.

Shen, X.; Lin, Z.; Brandt, J.; and Wu, Y. 2012. Mobile Product Image Search by Automatic Query Object Extraction. In *Proceedings of the European Conference on Computer Vision*, 114–127. Springer.

Singh, A. 2021. CLDA: Contrastive Learning for Semi-supervised Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 34, 5089–5101.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying Semi-supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems*, volume 33, 596–608.

Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Deep Spatial-semantic Attention for Fine-grained Sketch-based Image Retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, 5551–5560.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial Discriminative Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5018–5027.

Wang, X.; Hu, P.; Liu, P.; and Peng, D. 2022. Deep Semisupervised Class- and Correlation-Collapsed Cross-View Learning. *IEEE Transactions on Cybernetics*, 52(3): 1588–1601.

Wang, X.; Peng, D.; Hu, P.; and Sang, Y. 2019. Adversarial correlated autoencoder for unsupervised multi-view representation learning. *Knowledge-Based Systems*, 168: 109–120.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Wynne, G.; and Duncan, A. B. 2022. A Kernel Two-sample Test for Functional Data. *Journal of Machine Learning Research*, 23(73): 1–51.

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C.-C. 2016. Sketch Me That Shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 799–807.

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021. Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13834–13844.

Zhu, Y.; Zhuang, F.; and Wang, D. 2019. Aligning Domain-specific Distribution and Classifier for Cross-domain Classification from Multiple Sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5989–5996.