DiDA: Disambiguated Domain Alignment for Cross-Domain Retrieval with Partial Labels

Haoran Liu[1, 2], **Ying Ma**[3], **Ming Yan**[4], **Yingke Chen**[5], **Dezhong Peng**[1, 2], **Xu Wang**[1*]

[1]College of Computer Science, Sichuan University, Chengdu, China
[2]National Innovation Center for UHD Video Technology, Chengdu, China
[3]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[4]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[5]Department of Computer and Information Sciences, Northumbria University, UK
HrLiu.ai@gmail.com, y.ma@hit.edu.cn, yanmingtop@gmail.com, yke.chen@gmail.com, pengdz@scu.edu.cn, wangxu.scu@gmail.com

Abstract

Driven by generative AI and the Internet, there is an increasing availability of a wide variety of images, leading to the significant a nd p opular t ask o f c ross-domain i mage retrieval. To reduce annotation costs and increase performance, this paper focuses on an untouched but challenging problem, i.e., cross-domain image retrieval with partial labels (PCIR). Specifically, PCIR faces great challenges due to the ambiguous supervision signal and the domain gap. To address these challenges, we propose a novel method called disambiguated domain alignment (DiDA) for cross-domain retrieval with partial labels. In detail, DiDA elaborates a novel prototype-score unitization learning mechanism (PSUL) to extract common discriminative representations by simultaneously disambiguating the partial labels and narrowing the domain gap. Additionally, DiDA proposes a prototype-based domain alignment mechanism (PBDA) to further bridge the inherent cross-domain discrepancy. Attributed to PSUL and PBDA, our DiDA effectively excavates domain-invariant discrimination for cross-domain image retrieval. We demonstrate the effectiveness of DiDA through comprehensive experiments on three benchmarks, comparing it to existing state-of-the-art methods. Code available: https://github.com/lhrrrrrr/DiDA.

1 Introduction

With the proliferation of digital platforms and the continuous generation of visual content, the need to organize, search, and retrieve images effectively has become paramount. However, traditional image retrieval approaches often suffer from limitations when confront with different visual domains, e.g., medical images, art, fashion, and satellite imagery. This is where cross-domain image retrieval (CIR) has emerged as a promising research direction to overcome these challenges, facilitating the exploration of images across diverse domains. Given a query image from one domain, CIR aims to retrieve images from different domains based on the similarity of visual representations. Notably, CIR has important research implications in deepening the comprehension of transferable visual features, along with practical applications in many scenarios such as surveillance (Liu et al. 2019) and e-commerce platforms (Lei et al. 2021).

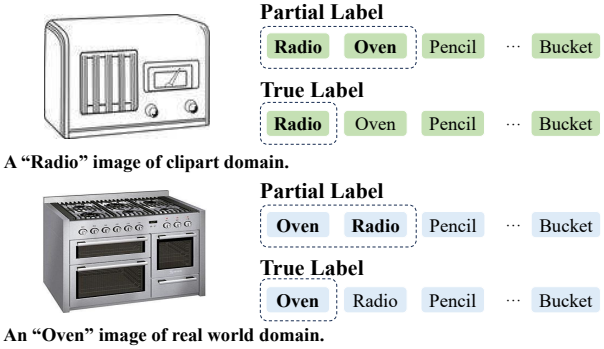


Figure 1: Two images from different domains are equipped with partial labels.

One of the major challenges faced by CIR is the domain gap caused by inconsistent feature distributions across distinct domains. To tackle this challenge, a rich line of studies (Sangkloy et al. 2016; Bhunia et al. 2022; Sain et al. 2023; Wang et al. 2022b; Sain et al. 2021; Wang et al. 2022c,a, 2023a) have been proposed. Among them, fully supervised learning for CIR achieves excellent performance attributed to its reliance on precise annotated labels. However, acquiring such precise annotations for diverse domains can be costly, time-consuming, and even require expert knowledge. Undoubtedly, these limitations hinder the scalability and practicality of existing CIR methods when dealing with large-scale datasets or evolving domains. To overcome the need for annotations, several unsupervised CIR (UCIR) approaches (Kim et al. 2021; Wang et al. 2023b) have been proposed. While these methods could learn latent relationships and semantics within and across domains without labels or correspondence, their retrieval performance is not yet promising.

To strike a balance between high labeling costs and achieving excellent performance, this paper introduces a new paradigm called cross-domain image retrieval with partial labels (PCIR). In PCIR, each sample is equipped with a

*Corresponding author.

set of candidate labels, with the true label hidden within this set. In other words, each training instance carries an annotation that contains ambiguity. This paradigm reflects real-world cross-domain data annotation scenarios where ambiguity exists in labeling tasks. For instance, in Fig. 1, the image labeled as "Radio" presents an inherent ambiguity that may make it challenging for the annotator to distinguish whether it belongs to the category of "Radio" or "Oven". Similarly, the image labeled as "Oven" from another domain also faces the same issue. Consequently, the annotator can consider both options as potential labels, resulting in partial labels. To the best of our knowledge, PCIR has not been explored in previous studies. Compared to CIR and UCIR, the challenge of PCIR lies in learning discriminative representations from partial labels with ambiguity while bridging the inherent gap across different domains.

To tackle the challenge, we propose a novel method termed DiDA. This method effectively unifies two mechanisms, namely the prototype-score unitization learning mechanism (PSUL) and the prototype-based domain alignment mechanism (PBDA), to accomplish the task of PCIR. Unlike existing approaches in partial label learning, PSUL enjoys the advantage of simultaneously achieving label disambiguation and mitigating domain gaps. To be specific, PSUL transforms the prototypes from different domains into prototype-scores and progressively brings them closer to the target unit matrix. This process not only bridges the prototypes from distinct domains but also facilitates the network to learn more accurate class probabilities, thus promoting label disambiguation. Additionally, to enhance domain alignment, PBDA computes the similarities between learned representations and prototypes from different domains. By reducing the discrepancy between these similarities, PBDA further narrows the cross-domain gap.

The main contributions are summarized as follows: (1) We propose a novel method called **Di**sambiguated **D**omain **A**lignment (**DiDA**) to tackle an untouched problem, i.e., cross-domain image retrieval with partial labels. To the best of our knowledge, this work could be the first study on this problem. (2) A novel prototype-score unitization learning mechanism (PSUL) is presented to learn common discrimination by alleviating the domain gap and promoting label disambiguation. (3) A novel prototype-based domain alignment mechanism (PBDA) is proposed to learn domain-invariant information and further eliminate the inherent discrepancy across different domains. (4) Extensive experiments are conducted on three widely-used benchmarks, showcasing the promising performance of DiDA for the PCIR task. Notably, our DiDA consistently delivers remarkably stable performance as the partial rate increases, outperforming existing state-of-the-art methods.

2 Related Work

2.1 Partial Label Learning

To learn the objective information from partial labels, numerous approaches are proposed to alleviate the ambiguities and improve the performance of the model. One typical strategy is to consider each candidate label equally while averaging the modeling outputs as prediction, named average-based methods (Hüllermeier and Beringer 2006; Cour, Sapp, and Taskar 2011; Zhang and Yu 2015). However, their performance is generally less effective than identification-based methods (Jin and Ghahramani 2002; Nguyen and Caruana 2008; Liu and Dietterich 2012; Yu and Zhang 2016). Identification-based methods treat the label as a latent variable and iteratively evolve the confidence of each candidate label. For example, Jin and Ghahramani (2002) adopt the maximum likelihood criterion and Yu and Zhang (2016) use the maximum margin criterion to identify the true label.

With the advances in deep neural networks, partial label learning (PLL) has drawn great attention. For instance, Feng et al. (2020) develop a risk-consistent method and a classifier-consistent method. Moreover, Wen et al. (2021) design a leveraged weighted loss to consider the impact of partial labels and non-partial labels simultaneously. Motivated by contrastive learning, PICO (Wang et al. 2021b) is presented to learn discriminative representations and employ the prototypes to disambiguate the partial label. Meanwhile, Wu, Wang, and Zhang (2022) rethink the utilization of consistency regularization and employ non-partial labels to perform supervised learning. After that, Xia et al. (2023) propose a guided prototypical classifier to facilitate the model to learn more effective representations. However, the aforementioned approaches are all implemented in one specific domain and the performance in cross-domain scenarios could be not satisfactory due to the huge domain gap. Therefore, PLL in cross-domain scenarios is still an unexplored and challenging issue.

2.2 Cross-domain Image Retrieval

Image retrieval is a fundamental task in computer vision that aims to retrieve relevant images from a large database based on a given query image. Cross-domain image retrieval (CIR) extends the conventional image retrieval task by addressing the challenge of searching for relevant images across different domains. To mitigate the domain gap and promote retrieval performance, numerous methods have been proposed. For example, Sangkloy et al. (2016) utilize both instance-level similarity and category-level similarity and Song et al. (2017) introduce a higher-order learnable energy function (HOLEF) based loss. Moreover, a transferable coupled network (Wang et al. 2021a) is presented for zero-shot sketch-based image retrieval. Motivated by meta-learning, Sain et al. (2021) propose a style-agnostic SBIR model which could dynamically adapt to unseen sketch styles. In addition, there are several approaches (Kim et al. 2021; Wang et al. 2023b) for solving unsupervised cross-domain retrieval (UCIR) tasks. For instance, Wang et al. (2023b) employ a correspondence-free domain alignment (CoDA) strategy to boost retrieval performance without correspondence and category annotations. In this paper, we focus on a new paradigm, i.e., cross-domain image retrieval with partial labels (PCIR). To the best of our knowledge, PCIR has not been touched in previous studies. Compared with CIR and UCIR, PCIR involves learning domain-invariant information from ambiguous annotation information, which is challenging but meaningful.

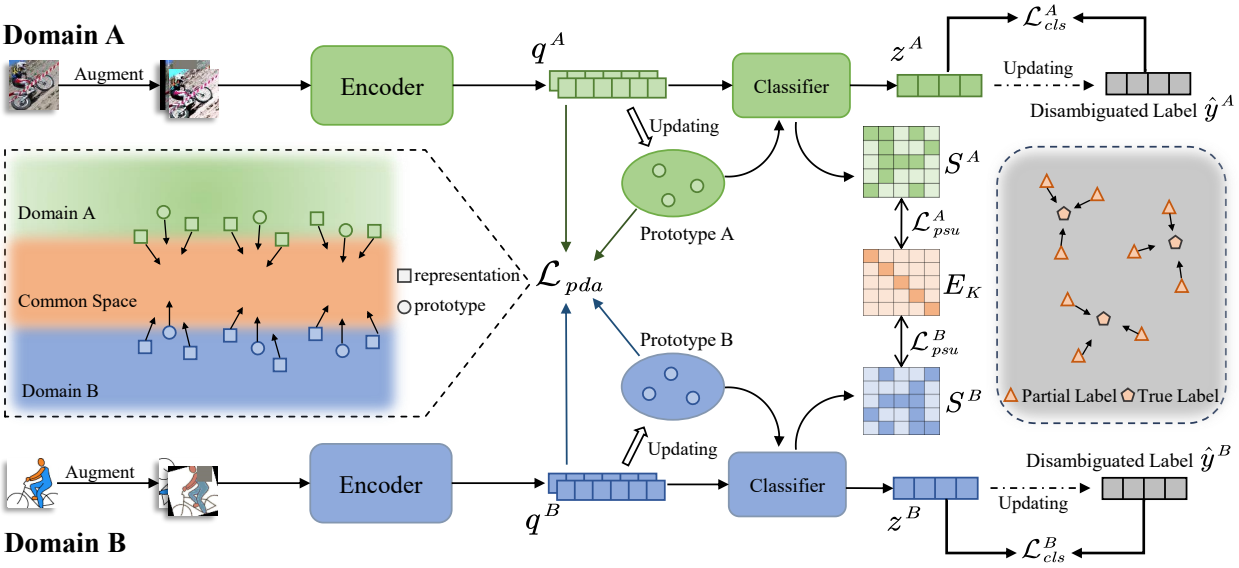


Figure 2: The pipeline of our DiDA for cross-domain image retrieval with partial labels. PSUL ($\mathcal{L}_{cls}$ and $\mathcal{L}_{psu}$) facilitates the model to extract discriminative and domain-invariant representations while promoting label disambiguation. Meanwhile, PBDA ($\mathcal{L}_{pda}$) further bridges the inherent gap across different domains.

3 Methodology

3.1 Problem Formulation

Notations. For a clear presentation, we first give the formal definition of the cross-domain retrieval task with partial labels (PCIR). Let $\mathcal{X}$ be the input space, $\mathcal{Y} = \{1, 2, ..., K\}$ be the label space with K classes. Given two sets of training datasets $\mathcal{D}_A = \{(\boldsymbol{x}_i^A, Y_i^A)\}_{i=1}^{N_A}$ from domain A and $\mathcal{D}_B = \{(\boldsymbol{x}_i^B, Y_i^B)\}_{i=1}^{N_B}$ from domain B, PCIR assumes that each input image $\boldsymbol{x} \in \mathcal{X}$ has a candidate label set $Y \subseteq \mathcal{Y}$. We only know the true label is included in Y but it's not clear which one is. Further, we define the vector form of Y as $\boldsymbol{y} \in \mathbb{R}^K$, where the element corresponding to the class in Y is 1 and the rest are 0. Our goal is to learn a feature encoder $f(\cdot)$ which is able to learn discriminative representation $\boldsymbol{q}$ from partial labels and efficiently bridge the domain gap. Formally, given a data point $\boldsymbol{x}$, the embedded features can be computed by $\boldsymbol{q} = f(\boldsymbol{x}) \in \mathbb{R}^L$, where L is the dimension of the common space. Meanwhile, we define a softmax classifier $g(\cdot)$ to transform the representation $\boldsymbol{q}$ into the probability distribution $\boldsymbol{z} = g(\boldsymbol{q}) \in \mathbb{R}^K$. Additionally, to enhance the generalization and robustness of the network, we construct weak augmentation $Aug_w(\cdot)$ and strong augmentation $Aug_s(\cdot)$. The representations of all the augmentations from domain A are denoted as $\mathcal{Q}^A = \{\mathcal{Q}_w^A \cup \mathcal{Q}_s^A\}$, where $\mathcal{Q}_w^A = \{f(Aug_w(\boldsymbol{x}_i^A))\}_{i=1}^{N_A}$ and $\mathcal{Q}_s^A = \{f(Aug_s(\boldsymbol{x}_i^A))\}_{i=1}^{N_A}$. Meanwhile, the probability distributions of all the augmentations are $\mathcal{Z}^A = \{\mathcal{Z}_w^A \cup \mathcal{Z}_s^A\}$. Similarly, we have $\mathcal{Q}^B$ and $\mathcal{Z}^B$ for domain B.

Overview. To disambiguate partial labels and bridge the domain gap, a novel cross-domain image retrieval method (DiDA) is proposed to learn common discrimination from partial labels while aligning the domains. As shown in Fig. 2, we present a prototype-score unitization learning mechanism (PSUL) to learn domain-invariant features and promote label disambiguation. Furthermore, we propose a prototype-based domain alignment mechanism (PBDA) to further narrow the cross-domain gap. The overall objective function could be formulated as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{cls} + \alpha(t) \cdot \mathcal{L}_{psu}}_{PSUL} + \underbrace{\beta(t) \cdot \mathcal{L}_{pda}}_{PBDA}, \tag{1}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{psu}$ are the objectives employed by the prototype-score unitization learning mechanism (PSUL) and $\mathcal{L}_{pda}$ is the objective adopted by the prototype-based domain alignment mechanism (PBDA). Both $\alpha(t)$ and $\beta(t)$ are dynamic trade-off parameters increasing with the epoch number t. To train the proposed DiDA, we minimize the loss function in a batch-by-batch manner by using a stochastic gradient descent optimizer. In the following sections, we will elaborate on the proposed DiDA approach.

3.2 Prototype-Score Unitization Learning

To excavate the discrimination and domain-invariant information from partial labels, we rethink the characteristics of the prototype and propose the prototype-score unitization learning mechanism, which can facilitate the model to encapsulate more discriminative representations and bridge the domain gap while achieving label disambiguation. In this section, we first describe the label disambiguation approach and then introduce the disambiguation-based prototype setting. Finally, the prototype-score unitization learning mechanism is proposed.

Label disambiguation. The core of label disambiguation is how to identify the ground truth from the ambiguous partial label $\boldsymbol{y}$ and then iteratively train the model relying on

the guidance of the disambiguated label $\hat{\boldsymbol{y}}$. To achieve label disambiguation, we adopt a classifier-based disambiguation strategy as follows:

$$\hat{\boldsymbol{y}}^A = \boldsymbol{z}^A \circ \boldsymbol{y}^A,\ \hat{\boldsymbol{y}}^B = \boldsymbol{z}^B \circ \boldsymbol{y}^B, \tag{2}$$

where $\boldsymbol{z}^A \in \mathcal{Z}_w^A$ and $\boldsymbol{z}^B \in \mathcal{Z}_w^B$. The element in $\boldsymbol{z}$ can be considered as the probability that sample $\boldsymbol{x}$ is the corresponding class, so we employ the Hadamard product $\circ$ of the probability distribution $\boldsymbol{z}$ and the partial label $\boldsymbol{y}$ as the new disambiguated label $\hat{\boldsymbol{y}}$. Meanwhile, the new disambiguated label $\hat{\boldsymbol{y}}$ is normalized to the probability distribution.

To further promote label disambiguation and enhance the performance, we adopt a classification loss as follows:

$$\mathcal{L}_{cls}^A = \sum\nolimits_{\boldsymbol{z}^A \in \mathcal{Z}^A} D_{KL}(\boldsymbol{z}^A || \hat{\boldsymbol{y}}^A), \tag{3}$$

where D_{KL} is the Kullback-Leibler (KL) divergence which is used to quantify the difference between two probability distributions. Specifically, $\mathcal{L}_{cls}^A$ measures the KL divergence of $\boldsymbol{z}^A$ and $\hat{\boldsymbol{y}}^A$. At the beginning of training, we use normalized $\boldsymbol{y}$ as the initial $\hat{\boldsymbol{y}}$. Similarly, we have the classification loss function $\mathcal{L}_{cls}^B$ for domain B as follows:

$$\mathcal{L}_{cls}^B = \sum\nolimits_{\boldsymbol{z}^B \in \mathcal{Z}^B} D_{KL}(\boldsymbol{z}^B || \hat{\boldsymbol{y}}^B). \tag{4}$$

Finally, the total classification loss $\mathcal{L}_{cls}$ can be written as:

$$\mathcal{L}_{cls} = \mathcal{L}_{cls}^A + \mathcal{L}_{cls}^B. \tag{5}$$

Disambiguation-based prototype. Prototypes serve as compact representations of classes in a suitable embedding space, allowing the model to capture the essential characteristics of each class and enable efficient comparison and classification of new samples. To begin with, we describe the definition of the prototypes:

$$\boldsymbol{P}^A = \begin{bmatrix} \boldsymbol{p}_1^A & \boldsymbol{p}_2^A & \cdots & \boldsymbol{p}_K^A \end{bmatrix}, \tag{6}$$

$$\boldsymbol{P}^B = \begin{bmatrix} \boldsymbol{p}_1^B & \boldsymbol{p}_2^B & \cdots & \boldsymbol{p}_K^B \end{bmatrix}, \tag{7}$$

where $\boldsymbol{P}^A \in \mathbb{R}^{L \times K}$ and $\boldsymbol{P}^B \in \mathbb{R}^{L \times K}$ represent the prototypes of domain A and domain B respectively and $\boldsymbol{p}_k \in \mathbb{R}^L$ denotes the prototype of the corresponding class k. Meanwhile, both $\boldsymbol{P}^A$ and $\boldsymbol{P}^B$ are initialized with all zeros. In order to better evolve the prototypes, we use pseudo labels $\hat{k}^A$ and $\hat{k}^B$ to respectively select the most probable class for the samples $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$:

$$\hat{k}^A = argmax\ (\boldsymbol{z}^A \circ \boldsymbol{y}^A), \tag{8}$$

$$\hat{k}^B = argmax\ (\boldsymbol{z}^B \circ \boldsymbol{y}^B). \tag{9}$$

Specifically, the class with the largest probability in the partial label is chosen as pseudo label $\hat{k}$ at every epoch. Guided by the pseudo labels $\hat{k}^A$ and $\hat{k}^B$, we adopt a moving-average mechanism to stably update the class prototypes $\boldsymbol{p}_k^A$ and $\boldsymbol{p}_k^B$ with the normalized representations $\boldsymbol{q}^A$ and $\boldsymbol{q}^B$:

$$\boldsymbol{p}_k^A = \lambda(t)\boldsymbol{p}_k^A + (1-\lambda(t))\boldsymbol{q}^A, \quad \text{if } \hat{k}^A = k, \tag{10}$$

$$\boldsymbol{p}_k^B = \lambda(t)\boldsymbol{p}_k^B + (1-\lambda(t))\boldsymbol{q}^B, \quad \text{if } \hat{k}^B = k, \tag{11}$$

where $\boldsymbol{q}^A \in \mathcal{Q}_w^A$, $\boldsymbol{q}^B \in \mathcal{Q}_w^B$, $\lambda(t)$ is a dynamic momentum parameter and the prototypes $\boldsymbol{p}_k^A$ and $\boldsymbol{p}_k^B$ are further normalized. Intuitively, as the training progresses, the accuracy of identifying the pseudo label $\hat{k}$ is increasing and the learned representation $\boldsymbol{q}$ is more discriminative. Therefore, we set $\lambda(t) \in [0.9, 0.5]$ to gradually decrease with the epoch t.

Prototype-score unitization. As aforementioned, we reconsider the essential properties of the prototypes. For each class k, there is a specific prototype $\boldsymbol{p}_k$ to represent the center of the class. We argue that since the class k of each prototype $\boldsymbol{p}_k$ is certain, we can utilize it as supervisory information of the prototype $\boldsymbol{p}_k$ for network training, which is referred to as prototype-score unitization learning. Specifically, we transfer the prototype $\boldsymbol{p}_k$ to the prototype-score $\boldsymbol{s}_k$ by the classifier $g(\cdot)$ and perform fully supervised learning with the target unit vector $\boldsymbol{e}_k$. The prototype-score matrices $\boldsymbol{S}^A \in \mathbb{R}^{K \times K}$ and $\boldsymbol{S}^B \in \mathbb{R}^{K \times K}$ are defined as:

$$\boldsymbol{S}^A = g(\boldsymbol{P}^A) = \begin{bmatrix} \boldsymbol{s}_1^A & \boldsymbol{s}_2^A & \cdots & \boldsymbol{s}_K^A \end{bmatrix}, \tag{12}$$

$$\boldsymbol{S}^B = g(\boldsymbol{P}^B) = \begin{bmatrix} \boldsymbol{s}_1^B & \boldsymbol{s}_2^B & \cdots & \boldsymbol{s}_K^B \end{bmatrix}, \tag{13}$$

where $\boldsymbol{s}_k^A = g(\boldsymbol{p}_k^A) \in \mathbb{R}^K$ and $\boldsymbol{s}_k^B = g(\boldsymbol{p}_k^B) \in \mathbb{R}^K$ are the prototype-score of the k-th prototype of domain A and domain B respectively. We further define the target unit matrix of prototype-score matrices $\boldsymbol{S}^A$ and $\boldsymbol{S}^B$ as $\boldsymbol{E}_K \in \mathbb{R}^{K \times K}$:

$$\boldsymbol{E}_K = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \cdots & \boldsymbol{e}_K \end{bmatrix}, \tag{14}$$

where $\boldsymbol{e}_k$ serves as the corresponding target unit vector of $\boldsymbol{s}_k^A$ and $\boldsymbol{s}_k^B$ and $\boldsymbol{e}_k \in \mathbb{R}^K$ represents a one-hot vector with 1 for the k-th class and 0 for other classes. Obviously, our purpose is to make the prototype-score matrices $\boldsymbol{S}^A$ and $\boldsymbol{S}^B$ approximate the target unit matrix $\boldsymbol{E}_K$. Therefore, we design a loss function $\mathcal{L}_{psu}^A$ to unitize the prototype-score matrix $\boldsymbol{S}^A$ to $\boldsymbol{E}_K$:

$$\mathcal{L}_{psu}^A = \sum\nolimits_{k \in \mathcal{Y}} U(\boldsymbol{e}_k^A, \boldsymbol{s}_k^A), \tag{15}$$

where $U(\cdot, \cdot)$ denotes the cross-entropy loss. Similarly, we define the prototype-score unitization loss of domain B as:

$$\mathcal{L}_{psu}^B = \sum\nolimits_{k \in \mathcal{Y}} U(\boldsymbol{e}_k^B, \boldsymbol{s}_k^B). \tag{16}$$

Finally, the total prototype-score unitization loss can be written as:

$$\mathcal{L}_{psu} = \mathcal{L}_{psu}^A + \mathcal{L}_{psu}^B. \tag{17}$$

By minimizing Eq. (17), our model not only learns common discriminative representations but also boosts label disambiguation. Detailly, as the ambiguity of the labels decreases, the network could obtain more accurate pseudo labels to correctly update the prototypes for prototype-score unitization learning. Therefore, driven by $\mathcal{L}_{psu}$, the network can learn more accurate class probabilities, thus enhancing disambiguation. Obviously, promising positive feedback is formed. Moreover, the cross-domain discrepancy is alleviated as the prototype-score matrices of different domains simultaneously converge to the target unit matrix. Thanks to such a mechanism, $\mathcal{L}_{psu}$ endows our DiDA with the ability to learn domain-invariant features as well.

3.3 Prototype-Based Domain Alignment

With prototype-score unitization learning, the model learns discriminative representations while mitigating the domain gap. However, the discrepancy between different domains still exists and needs to be eliminated. Hence, it is necessary to further align the distinct domains and excavate

Method		OfficeHome			Office31			ImageCLEF		
		0.1	0.2	0.3	0.1	0.2	0.3	0.3	0.4	0.5
Supervised	best	0.733	0.733	0.733	0.940	0.940	0.940	0.787	0.787	0.787
	last	0.714	0.714	0.714	0.934	0.934	0.934	0.762	0.762	0.762
CDS (Kim et al. 2021)	best	0.452	0.452	0.452	0.738	0.738	0.738	0.709	0.709	0.709
	last	0.414	0.414	0.414	0.663	0.663	0.663	0.591	0.591	0.591
CoDA (Wang et al. 2023b)	best	0.486	0.486	0.486	0.788	0.788	0.788	0.727	0.727	0.727
	last	0.482	0.482	0.482	0.781	0.781	0.781	0.721	0.721	0.721
CC (Feng et al. 2020)	best	0.639	0.505	0.418	0.911	0.839	0.748	0.766	0.747	0.716
	last	0.576	0.499	0.410	0.891	0.811	0.703	0.743	0.730	0.703
RC (Feng et al. 2020)	best	0.641	0.505	0.418	0.910	0.841	0.746	0.764	0.744	0.710
	last	0.602	0.499	0.411	0.894	0.813	0.706	0.740	0.726	0.696
PRODEN (Lv et al. 2020)	best	0.645	0.505	0.419	0.911	0.840	0.746	0.764	0.742	0.710
	last	0.607	0.501	0.414	0.897	0.817	0.710	0.744	0.729	0.699
LWC (Wen et al. 2021)	best	0.642	0.503	0.418	0.905	0.834	0.742	0.763	0.743	0.711
	last	0.605	0.498	0.410	0.893	0.806	0.704	0.738	0.722	0.690
PICO (Wang et al. 2021b)	best	0.688	0.590	0.343	0.921	0.896	0.817	0.741	0.734	0.721
	last	0.673	0.579	0.268	0.915	0.893	0.809	0.721	0.727	0.717
DPLL (Wu, Wang, and Zhang 2022)	best	0.676	0.556	0.473	0.924	0.879	0.814	0.771	0.753	0.725
	last	0.626	0.550	0.467	0.912	0.858	0.784	0.719	0.721	0.694
PaPi (Xia et al. 2023)	best	0.662	0.525	0.437	0.923	0.866	0.792	0.750	0.724	0.691
	last	0.647	0.502	0.404	0.883	0.803	0.699	0.668	0.632	0.595
DiDA (Ours)	best	**0.712**	**0.653**	**0.594**	**0.930**	**0.911**	**0.890**	**0.781**	**0.771**	**0.764**
	last	**0.704**	**0.650**	**0.591**	**0.924**	**0.905**	**0.886**	**0.765**	**0.762**	**0.755**

Table 1: The average mAP retrieval performance comparison for our DiDA and other compared methods on OfficeHome, Office31 and ImageCLEF datasets under different partial rates (0.1-0.5). The best performance results are shown in bold.

domain-invariant information. For this purpose, we adopt a prototype-based domain alignment loss:

$$\mathcal{L}_{pda}^{A} = \sum_{\boldsymbol{q}^A \in \mathcal{Q}^A} \sum_{k \in \mathcal{Y}} |\boldsymbol{q}^A \cdot \boldsymbol{p}_k^A - \boldsymbol{q}^A \cdot \boldsymbol{p}_k^B|. \tag{18}$$

Specifically, we compute the similarities of the normalized representation $\boldsymbol{q}^A$ and the prototypes $\boldsymbol{p}_k^A$ and $\boldsymbol{p}_k^B$ respectively, and then employ the MAE loss to minimize the discrepancy between them. Meanwhile, we adopt both weakly augmented and strongly augmented images as inputs to encourage robustness and generalization. Similarly, we have the prototype-based domain alignment loss for domain B:

$$\mathcal{L}_{pda}^{B} = \sum_{\boldsymbol{q}^B \in \mathcal{Q}^B} \sum_{k \in \mathcal{Y}} |\boldsymbol{q}^B \cdot \boldsymbol{p}_k^B - \boldsymbol{q}^B \cdot \boldsymbol{p}_k^A|. \tag{19}$$

Finally, the total prototype-based domain alignment loss can be written as follows:

$$\mathcal{L}_{pda} = \mathcal{L}_{pda}^{A} + \mathcal{L}_{pda}^{B}. \tag{20}$$

As the $\mathcal{L}_{pda}$ decreases, the model focuses more on the common information across different domains. Both the representations from different domains and the domain-specific prototypes gradually converge to the common space.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our method, we conduct extensive comparison experiments on three cross-domain benchmark datasets, i.e., Office31 (Saenko et al. 2010), OfficeHome (Venkateswara et al. 2017) and ImageCLEF (Long et al. 2017). These datasets are briefly introduced as follows: **Office31:** Office31 is a benchmark dataset with three object domains: Amazon (A), DSLR (D) and Webcam (W). The dataset consists of 31 categories of images and the three domains contain 2817, 498, and 795 images respectively. We conduct six retrieval tasks, i.e., A-D, A-W, D-A, D-W, W-A and W-D. **OfficeHome:** OfficeHome is a large dataset of 15,500 images which has 65 categories and four domains: Artistic (A), Clipart (C), Product (P), and Real-world (R). **Image-CLEF:** The Image-CLEF dataset is composed of four domains: Bing (B), Caltech256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). It has 12 categories and each domain has 600 images. We perform twelve retrieval tasks on OfficeHome and Image-CLEF respectively. The above datasets are randomly partitioned into training sets and testing sets in an 80-20 ratio.

4.2 Implementation Detail

In DiDA, we utilize the ResNet-50 network as the encoder and initialize it with parameters pre-trained in ImageNet. Note that, the last fully connected layer is substituted by a 512-D randomly initialized linear layer and the output features are l_2-normalized. Meanwhile, the classifier consists of a linear layer and is initialized by the Xavier initialization method (Glorot and Bengio 2010). Furthermore, we adopt the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and set the learning rate to 0.003 and 0.0001 for the encoder and classifier respectively. For a fair compar-

Method	Cross-domain Retrieval Task on OfficeHome Dataset (Partial Rate:0.1)												
	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
Supervised	0.683	0.678	0.660	0.625	0.787	0.723	0.658	0.788	0.793	0.648	0.731	0.790	0.714
CDS	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
CoDA	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
CC	0.525	0.593	0.561	0.400	0.652	0.575	0.483	0.652	0.726	0.452	0.587	0.707	0.576
RC	0.526	0.617	0.585	0.438	0.685	0.608	0.520	0.674	0.738	0.491	0.604	0.737	0.602
PRODEN	0.536	0.609	0.583	0.439	0.692	0.615	0.531	0.682	0.752	0.494	0.617	0.734	0.607
LWC	0.541	0.611	0.588	0.432	0.680	0.608	0.525	0.680	0.751	0.496	0.605	0.737	0.605
PICO	0.579	0.652	0.639	0.573	0.746	0.678	0.630	0.735	0.778	0.604	0.678	0.782	0.673
DPLL	0.552	0.607	0.600	0.413	0.726	0.663	0.528	0.725	0.773	0.497	0.668	0.757	0.626
PaPi	0.563	0.623	0.613	0.498	0.722	0.681	0.585	0.706	0.777	0.563	0.662	0.769	0.647
DiDA (Ours)	**0.628**	**0.677**	**0.662**	**0.612**	**0.764**	**0.714**	**0.656**	**0.767**	**0.793**	**0.657**	**0.712**	**0.801**	**0.704**

Table 2: The mAP retrieval performance (obtained from the last epoch) comparison for our DiDA and other compared methods on the OfficeHome dataset under partial rate: 0.1. The best performance results are shown in bold.

ison with baselines, the batch size is set to 16 and the total epochs are 50. All the experiments are carried out using PyTorch with two Nvidia GeForce RTX 3090 GPUs.

4.3 Experimental Setup

To validate the effectiveness of our proposed method, we compare our DiDA with seven PLL methods, two unsupervised cross-domain methods and one fully supervised method with true labels. The compared PLL methods are as follows: **CC** and **RC** (Feng et al. 2020), **PRODEN** (Lv et al. 2020), **LWC** (Wen et al. 2021), **PICO** (Wang et al. 2021b), **DPLL** (Wu, Wang, and Zhang 2022) and **PaPI** (Xia et al. 2023). The fully supervised method is implemented by the cross-entropy loss with ground-truth labels. Additionally, we adopt **CDS** (Kim et al. 2021) and **CoDa** (Wang et al. 2023b) as the unsupervised methods. For a fair comparison, all the methods utilize the same 512-D features from the encoder for retrieval. Moreover, our evaluation metric employs mean average precision (mAP) on all retrieved results and we report the mAP results for all methods. To more rigorously evaluate the robustness of the methods under various label ambiguities, we set three different partial rates for each dataset based on the number of categories of the datasets, i.e., $\{0.1, 0.2, 0.3\}$ for OfficeHome and Office31, $\{0.3, 0.4, 0.5\}$ for Image-CLEF.

4.4 Comparison with State-of-the-Art Methods

We conduct cross-domain image retrieval with partial labels on the three datasets to evaluate the performance of our DiDA and the compared methods. The experimental results under different partial rates are reported in Tables 1 and 2, with additional results available in the Supplementary. As shown in these tables, our DiDA is superior to the existing methods on the three datasets. From the experimental results, we could obtain the following observations: (1) As shown in Table 1, our DiDA outperforms other baselines on all datasets with different partial rates. For example, with the partial rate of 0.3, our DiDA exceeds DPLL by 0.121 on the OfficeHome dataset and PICO by 0.073 on the Office31 dataset. It demonstrates that DiDA more efficiently learns common discriminative features from partial labels while bridging the domain gap. (2) Since PLL methods (PICO, DPLL, PaPi, etc.) are designed for domain-specific tasks, they cannot achieve desirable results on multi-domain tasks. For example, they even underperform the unsupervised cross-domain retrieval methods when the partial rate is large. This reveals that the cross-domain discrepancy significantly impacts their performances. (3) In general, the performance of training with partial labels will be inferior to the fully supervised method. However, our method could achieve comparable even better performance to the fully supervised method when the partial rate is relatively low. For instance, as shown in Table 2, our DiDA surpasses the fully supervised method by 0.011 on the R-P retrieval task of the OfficeHome dataset, which shows the superiority of our approach. (4) Obviously, as the partial rate increases, the ambiguity of data labels also increases, resulting in a significant decline in the performance of the compared methods. However, our DiDA can against this increasing ambiguity more effectively. According to Table 1, as the partial rate increases from 0.1 to 0.3, the average mAP of our method reduces by only 0.118 on the OfficeHome dataset, while PICO, DPLL and PaPi decrease by 0.345, 0.203 and 0.225, respectively.

4.5 Ablation Study

In this section, we investigate the contribution of each proposed component (i.e., loss $\mathcal{L}_{cls}$, $\mathcal{L}_{psu}$ and $\mathcal{L}_{pda}$) for cross-domain image retrieval with partial labels. For this purpose, we conduct three variants of our method, i.e., Variant 1: DiDA with $\mathcal{L}_{cls}$ only; Variant 2: DiDA with $\mathcal{L}_{cls}$ and $\mathcal{L}_{psu}$; Variant 3: DiDA with $\mathcal{L}_{cls}$ and $\mathcal{L}_{pda}$. For a fair comparison, we perform ablation experiments on the OfficeHome and Office31 datasets under the same experimental settings. From the experimental results shown in Table 3, one can observe that $\mathcal{L}_{psu}$ can dramatically boost the performance, which indicates its effectiveness in label disambiguation and domain-gap reduction. Meanwhile, the results also demonstrate that $\mathcal{L}_{pda}$ further aligns the distinct domains. In con-

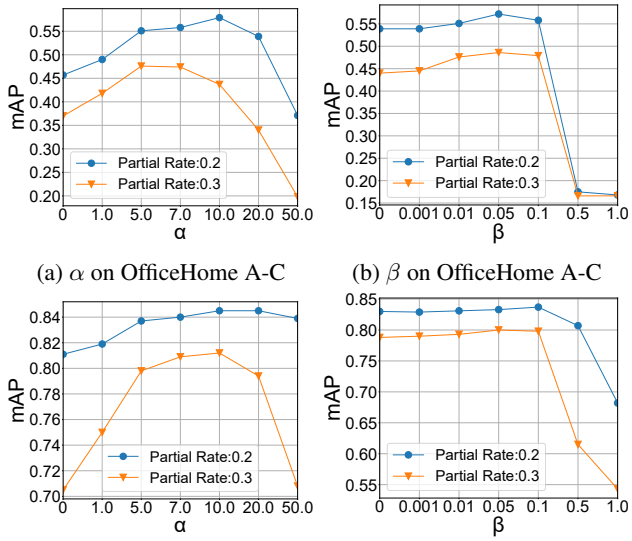
clusion, DiDA can effectively enhance the performance on distinct datasets under different partial rates.

Method	OfficeHome			Office31		
	0.1	0.2	0.3	0.1	0.2	0.3
Variant 1	0.673	0.560	0.477	0.919	0.861	0.788
Variant 2	0.698	0.644	0.573	0.923	0.904	0.868
Variant 3	0.698	0.583	0.503	0.923	0.893	0.828
Full DiDA	**0.704**	**0.650**	**0.591**	**0.924**	**0.905**	**0.886**

Table 3: The average mAP retrieval performance (obtained from the last epoch) comparison for the DiDA and its three variants on OfficeHome and Office31 datasets.

4.6 Effect of Coefficient α and β

To analyze the impact of the coefficient α and β in Eq. (1), we conduct parameter analysis experiments on the OfficeHome A-C task and Office31 D-A task. As shown in Fig. 3, we plot mAP scores w.r.t. different parameters of α and β. From the figure, we can observe that the model yields a stable performance when α and β are in two relatively large ranges (i.e., $\alpha \in [5, 10]$ and $\beta \in [0.01, 0.1]$) respectively.



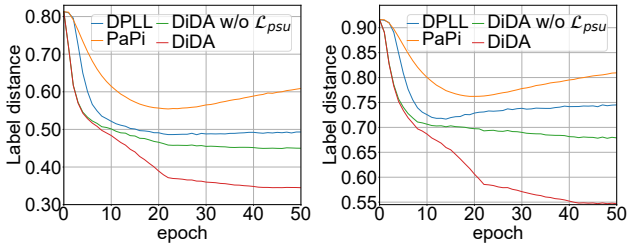
(a) α on OfficeHome A-C (b) β on OfficeHome A-C

(c) α on Office31 D-A (d) β on Office31 D-A

Figure 3: PCIR performance of DiDA in terms of mAP scores versus different values of α and β on OfficeHome A-C task and Office31 D-A task.

4.7 Performance of Label Disambiguation

To visually investigate the performance of label disambiguation, we plot the label distance (i.e., the Euclidean distance between the disambiguated label and the true label) versus epochs for our DiDA, DiDA without $\mathcal{L}_{psu}$ and the compared methods (i.e., DPLL and PaPi). As shown in Fig. 4, we conduct the experiments on the Office31 A-D task under partial rates of 0.1 and 0.2. It is evident that the proposed method excels in disambiguation performance and $\mathcal{L}_{psu}$ makes a significant contribution to disambiguation. This showcases the effectiveness of DiDA in eliminating ambiguity from partial labels and approaching ground truth within the label space.

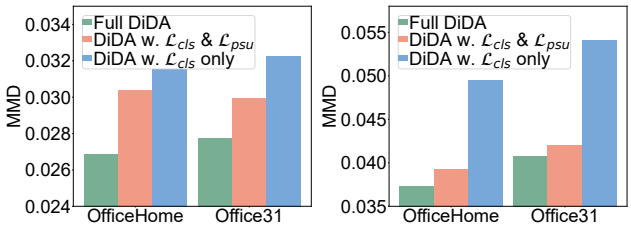


(a) Partial Rate: 0.1 (b) Partial Rate: 0.2

Figure 4: Label disambiguation performance versus epochs on Office31 A-D task under different partial rates.

4.8 Effect of Domain-Gap Elimination

To investigate the contribution of $\mathcal{L}_{psu}$ and $\mathcal{L}_{pda}$ on domain-gap elimination, we plot the domain discrepancy in terms of maximum mean discrepancy (MMD) for our DiDA, DiDA w. $\mathcal{L}_{cls}$ & $\mathcal{L}_{psu}$ and DiDA w. $\mathcal{L}_{cls}$ only under different partial rates. As shown in Fig. 5, we conduct the experiments on the OfficeHome A-C task and Office31 D-A task. It demonstrates that $\mathcal{L}_{psu}$ can effectively achieve domain-gap reduction and $\mathcal{L}_{pda}$ can further minimize the domain discrepancy.



(a) Partial Rate: 0.1 (b) Partial Rate: 0.2

Figure 5: The MMD of distinct domains on Office31 D-A task and OfficeHome A-C task under different partial rates.

5 Conclusion

In this paper, we study a new problem, i.e., cross-domain image retrieval with partial labels (PCIR). To this end, a novel method termed DiDA is proposed to project distinct domains into a common space under the supervision of partial labels. Specifically, our DiDA adopts a novel prototype-score unitization learning mechanism (PSUL) to encapsulate discriminative features into the domain-invariant space while achieving label disambiguation. Meanwhile, we employ a novel prototype-based domain alignment mechanism (PBDA) to eliminate the inherent gap across different domains further. Comprehensive experiments are conducted compared to several state-of-the-art approaches on three multi-domain datasets, demonstrating the effectiveness of our DiDA.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U19A2078, 62306197, 62372315), China Postdoctoral Science Foundation (2021TQ0223, 2022M712236), Sichuan Science and Technology Planning Project (2023YFG0033, 2023ZHCG0016, 2023YFQ0020, 2023ZYD0143), Chengdu Science and Technology Project (2023-XT00-00004-GX, 2021-JB00-00025-GX), Postdoctoral Joint Training Program of Sichuan University (SCDXLHPY2307).

References

Bhunia, A. K.; Koley, S.; Khilji, A. F. U. R.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 999–1008.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *The Journal of Machine Learning Research*, 12: 1501–1536.

Feng, L.; Lv, J.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably consistent partial-label learning. *Advances in neural information processing systems*, 33: 10948–10960.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Hüllermeier, E.; and Beringer, J. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5): 419–439.

Jin, R.; and Ghahramani, Z. 2002. Learning with multiple labels. *Advances in neural information processing systems*, 15.

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.

Lei, H.; Chen, S.; Wang, M.; He, X.; Jia, W.; and Li, S. 2021. A new algorithm for sketch-based fashion image retrieval based on cross-domain transformation. *Wireless Communications and Mobile Computing*, 2021: 1–14.

Liu, J.; Zha, Z.-J.; Chen, D.; Hong, R.; and Wang, M. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7202–7211.

Liu, L.; and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. *Advances in neural information processing systems*, 25.

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.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive identification of true labels for partial-label learning. In *international conference on machine learning*, 6500–6510. PMLR.

Nguyen, N.; and Caruana, R. 2008. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 551–559.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 213–226. Springer.

Sain, A.; Bhunia, A. K.; Chowdhury, P. N.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2765–2775.

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.

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.

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.

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, H.; Deng, C.; Liu, T.; and Tao, D. 2021a. Transferable coupled network for zero-shot sketch-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9181–9194.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2021b. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*.

Wang, Q.; Tao, Z.; Gao, Q.; and Jiao, L. 2022a. Multi-View Subspace Clustering via Structured Multi-Pathway Network. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Q.; Tao, Z.; Xia, W.; Gao, Q.; Cao, X.; and Jiao, L. 2022b. Adversarial multiview clustering networks with adaptive fusion. *IEEE transactions on neural networks and learning systems*.

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

Wang, X.; Peng, D.; Hu, P.; Gong, Y.; and Chen, Y. 2023a. Cross-Domain Alignment for Zero-Shot Sketch-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, X.; Peng, D.; Yan, M.; and Hu, P. 2023b. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10200–10208.

Wen, H.; Cui, J.; Hang, H.; Liu, J.; Wang, Y.; and Lin, Z. 2021. Leveraged weighted loss for partial label learning. In *International Conference on Machine Learning*, 11091–11100. PMLR.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2022. Revisiting consistency regularization for deep partial label learning. In *International Conference on Machine Learning*, 24212–24225. PMLR.

Xia, S.; Lv, J.; Xu, N.; Niu, G.; and Geng, X. 2023. Towards Effective Visual Representations for Partial-Label Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15589–15598.

Yu, F.; and Zhang, M.-L. 2016. Maximum margin partial label learning. In *Asian conference on machine learning*, 96–111. PMLR.

Zhang, M.-L.; and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, 4048–4054.