This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unsupervised Feature Representation Learning for Domain-generalized Cross-domain Image Retrieval

Conghui Hu Can Zhang Gim Hee Lee
Department of Computer Science, National University of Singapore
conghui@nus.edu.sg can.zhang@u.nus.edu gimhee.lee@nus.edu.sg

Abstract

Cross-domain image retrieval has been extensively studied due to its high practical value. In recently proposed unsupervised cross-domain image retrieval methods, efforts are taken to break the data annotation barrier. However, applicability of the model is still confined to domains seen during training. This limitation motivates us to present the first attempt at domain-generalized unsupervised cross-domain image retrieval (DG-UCDIR) aiming at facilitating image retrieval between any two unseen domains in an unsupervised way. To improve domain generalizability of the model, we thus propose a new two-stage domain augmentation technique for diversified training data generation. DG-UCDIR also shares all the challenges present in the unsupervised cross-domain image retrieval, where domain-agnostic and semantic-aware feature representations are supposed to be learned without external supervision. To accomplish this, we introduce a novel cross-domain contrastive learning strategy by utilizing phase image as a proxy to mitigate the domain gap. Extensive experiments are carried out using PACS and DomainNet dataset, and consistently illustrate the superior performance of our framework compared to existing state-of-the-art methods. Our source code is available at https://github.com/conghui1002/DG-UCDIR.

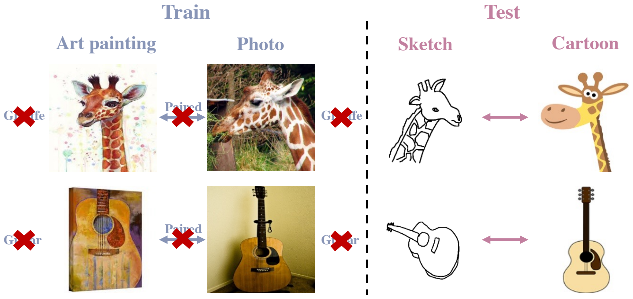


Figure 1. Illustration of DG-UCDIR. Given only unlabeled training domain data, the objective of DG-UCDIR is to facilitate retrieval between unseen domains.

1. Introduction

Cross-domain image retrieval finds a variety of applications in online shopping [15, 31], law enforcement [18, 20], *etc*. Most existing cross-domain image retrieval algorithms [27, 32] heavily rely on category and pair annotations to drive explicit semantic-aware and domain-invariant feature learning for retrieval. One emerging research direction for reducing data labeling cost is unsupervised cross-domain image retrieval (UCDIR) [9, 10], where human annotation is no longer necessary for model training and the trained model can be employed to conduct retrieval between seen domains. However, even the raw image can be labor-intensive to acquire for some domains such as sketch. As reported in [6], the median drawing time for one non-expert sketch in the TU-Berlin dataset is 86 seconds. This largely motivates us to investigate the domain-generalized unsupervised cross-domain image retrieval (DG-UCDIR) that sidesteps both data labeling and collection barriers, as illustrated in Fig. 1. Specifically, our ultimate goal is to leverage only unlabeled data from a pair of seen domains, *e.g*., art painting and real photo, to train a feature extractor which is capable of projecting imagery input from any unseen domain, *e.g*., sketch and cartoon, into a domain invariant shared feature space. Images from one unseen domain, *e.g*., sketch, can then be used as queries to conduct retrieval of the same category data from another unseen domain, *e.g*., cartoon. Consequently, DG-UCDIR is particularly valuable for those domains with data scarcity bottleneck.

Nevertheless, DG-UCDIR is an extremely challenging task due to the requirements of: 1) **Novel domain generalizability.** Given data from only the seen domain, we aim to facilitate image retrieval across to unseen domains, *i.e*., endow trained model with domain generalizability. Both existing supervised [32] and unsupervised [9] cross-domain image retrieval methods are not suited for direct testing on the novel domains since they are designed for specific domain pairs and thus is highly susceptible to failure on unseen novel domains. 2) **Domain-agnostic unsupervised feature learning.** The absence of category label or pairwise supervision for the seen domains during training, makes it extremely difficult for the model to learn semantically meaningful and domain-invariant features for effective category-level retrieval. The task then boils down to domain-agnostic unsupervised feature representation learning. Although contrastive learning [2, 3] has shown great promise in the context of unsupervised learning, the vanilla contrastive learning algorithm neglects influence of domain-specific knowledge and thus still suffers from the inability to mitigate domain shift and features alignment from different domains.

In this paper, we introduce a novel deep learning framework that simultaneously performs unseen domain generalization and domain-agnostic feature learning in an unsupervised paradigm. To address the first challenge of DG-UCDIR, we propose a new domain augmentation strategy by exploiting inherent characteristics of frequency domain data. More concretely, Fourier transform is employed to convert the RGB image into the disentangled phase and amplitude component in the frequency domain. According to [29] and [30], class-discriminative knowledge is mainly contained in the high-frequency portion of the phase part. We thus design a two-stage domain augmentation technique by first augmenting the low-frequency phase component, followed by the amplitude component distortion of the original image, while keeping the semantic information useful for image retrieval unchanged. As a result, our framework can be more resilient to the overfitting on seen domains during training compared to existing cross-domain image retrieval methods. For domain-agnostic unsupervised feature learning, we devise a set of phase-enhanced contrastive losses to: 1) remedy shifts in the seen domains that would hinder the effective training of feature extractor; 2) further bridge the gap between seen and unseen domains for better generalization; 3) mitigate the discrepancy between unseen domains for more effective cross-domain image retrieval. Intuitively, we leverage the smaller domain gap between phase images compared to their RGB counterparts (*c.f*. Fig. 2) to conduct unsupervised feature learning. Specifically, we formulate: 1) a phase-enhanced ***instance-instance*** contrastive loss by regarding the phase image as the positive pair of the corresponding RGB input, *i.e*., the phase image is adopted as a proxy to ameliorate the domain discrepancy and align features from various domains; 2) a phase-enhanced ***instance-centroid*** contrastive loss, where centroids shared by different domains are measured based on the phase image features to help pull semantically similar instance closer across domains.

Our main contributions are summarized as follows:

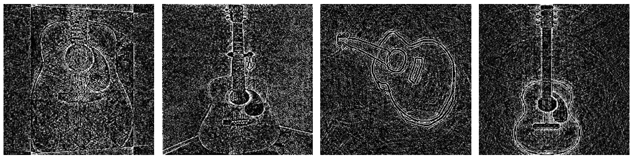


Figure 2. Smaller domain gaps are seen in the phase images compared to the RGB guitar images from four domains in Figure 1.

1. We introduce a new research direction of domain-generalized unsupervised cross-domain image retrieval (DG-UCDIR) which aims at advancing the practical application of image retrieval.
2. A two-stage domain augmentation strategy is proposed to increase the data diversity for superior novel domain generalizability.
3. We design the phase-enhanced instance-instance and instance-centroid contrastive losses to effectively facilitate the unsupervised domain-invariant and semantic-aware feature learning.
4. Extensive experiments on PACS and DomainNet validate the efficacy of our proposed framework on DG-UCDIR.

2. Related Work

Cross-domain Image Retrieval. Cross-domain image retrieval [11] targets at retrieving images from a target domain using the query image from another source domain. In the context of category-level cross-domain image retrieval [24, 9], only images of the same category as the query image are correct retrievals, which requires semantically meaningful features to be aligned across domains. For supervised cross-domain image retrieval, human-annotated class label [25] and data pairs [24, 31] can be employed to learn a shared feature space where valid cross-domain feature distance can be measured for effective retrieval. To circumvent the tedious data labeling process, unsupervised cross-domain image retrieval [9, 10] is therefore proposed. Provided with only unlabelled data, [9] introduces a distance-of-distance loss to minimize the domain discrepancy, while [10] leverages optimal transport [22] to estimate the cross-domain correspondence. However, it is still time-consuming to collect enough unlabeled images to facilitate model training for domains like sketch. We are therefore inspired to explore the domain-generalized unsupervised cross-domain image retrieval task to further break the data collection barrier.

Contrastive Learning. Contrastive learning [26, 2, 28] is a popular self-supervised learning method where the feature representation is learned by contrasting positive and negative pairs. Vanilla contrastive learning methods are proposed for single-domain data. [2] and [1] propose to pull different variants of the same sample closer in feature-level and cluster assignment-level, respectively. Feature memory bank is employed in [3] to aggregate features from the previous batch for more effective training. In addition, to extract meaningful semantic information, prototypes are introduced in [14]. More recently, to deal with multi-domain data, [33] explores the efficacy of cross-domain instance-prototype contrastive loss which can align the features of domains involved in training. Nevertheless, existing research on multi-domain contrastive learning still neglects the domain gap between seen and unseen domains in DG-UCDIR, and the feature space learned with training domains is not directly applicable to novel domains.

Domain Generalization. According to the number of training source domains, domain generalization techniques can be categorized into: 1) Multi-source domain generalization [30, 16] where a set of source domains are simultaneously employed for training to endow the trained model with generalization ability; 2) Single-source domain generalization [5, 4] which is a more challenging task as only the knowledge from one single domain can be leveraged to enable the novel domain test. Moreover, unsupervised domain generalization [7] is another newly proposed research direction aiming at removing the data labeling burden and training the classification model with only raw data without annotation. In terms of application, our DG-UCDIR is different from the original DG and attempts to facilitate image retrieval between novel domains provided with a pair of unlabeled seen domains.

3. Our Methodology

Overview. In DG-UCDIR, our goal is to leverage only unlabeled seen domain data to learn a feature extractor f_θ which can be applied to extract unseen domain features for effective cross-domain image retrieval. Formally, the feature extractor f_θ trained with imagery data $\mathcal{I}^A = \left\{I_i^A\right\}_{i=1}^M$ and $\mathcal{I}^B = \left\{I_j^B\right\}_{j=1}^N$ from seen domain A and B in an unsupervised manner is used to embed pixel-level input $\mathcal{I}^C = \left\{I_i^C\right\}_{i=1}^P$ and $\mathcal{I}^D = \left\{I_j^D\right\}_{j=1}^Q$ from unseen domain C and D into the feature space. M, N, P and Q stand for the number of images in domain A, B, C and D respectively. $\mathbf{x}_i^C$ and $\mathbf{x}_j^D$ are feature representations for instances in domain C and D, respectively. For domain C → D image retrieval, feature distance d_f between the query feature $\mathbf{x}_i^C$ and features $\left\{\mathbf{x}_j^D\right\}_{j=1}^Q$ in domain D are used as the criterion to rank all images from domain D in ascending order. For the query image I_i^C of class S, the correct retrievals should be domain D images belonging to the same category S, and ideally, they should be ranked at the top of the retrieval list. It is apparent that the feature extractor f_θ needs to be able to generalize the knowledge acquired from domains A and B to novel domains C and D, *i.e.*, domain generalizability. In addition, there is no available external supervision such as class label and pair information during the whole training process, and thus aggravating the difficulty for f_θ to predict domain-invariant and semantic-aware features that make the cross-domain feature distance d_f meaningful. To address the identified issues of DG-UCDIR, we propose: 1) A **two-stage domain augmentation strategy** to generate more domains for training, with the aim of strengthening domain generalizability; 2) **Phase-enhanced contrastive learning losses** where phase image is introduced to facilitate domain-agnostic semantic structure encoding.

3.1. Two-stage Domain Augmentation

The underlying assumption for our two-stage domain augmentation is that the high-level semantic information mainly exists in the high-frequency phase components from the Fourier transformed frequency domain, and the distortion of low-frequency phase and amplitude components do not change the categorical features. The frequency representation F_i for image $I_i \in \mathbb{R}^{C\times H\times W}$ by applying Fourier transform $\mathcal{F}$ is:

$$\begin{aligned} F_i(u,v) &= \mathcal{F}(I_i)(u,v) \\ &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} I_i(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}. \end{aligned} \tag{1}$$

For brevity, channel dimension C is omitted here. The corresponding phase PH_i and amplitude AM_i components become:

$$\begin{aligned} \mathrm{PH}_i(u,v) &= \arctan\frac{\mathrm{Im}(F_i(u,v))}{\mathrm{Re}(F_i(u,v))}, \\ \mathrm{AM}_i(u,v) &= \sqrt{\mathrm{Im}(F_i(u,v))^2 + \mathrm{Re}(F_i(u,v))^2}, \end{aligned} \tag{2}$$

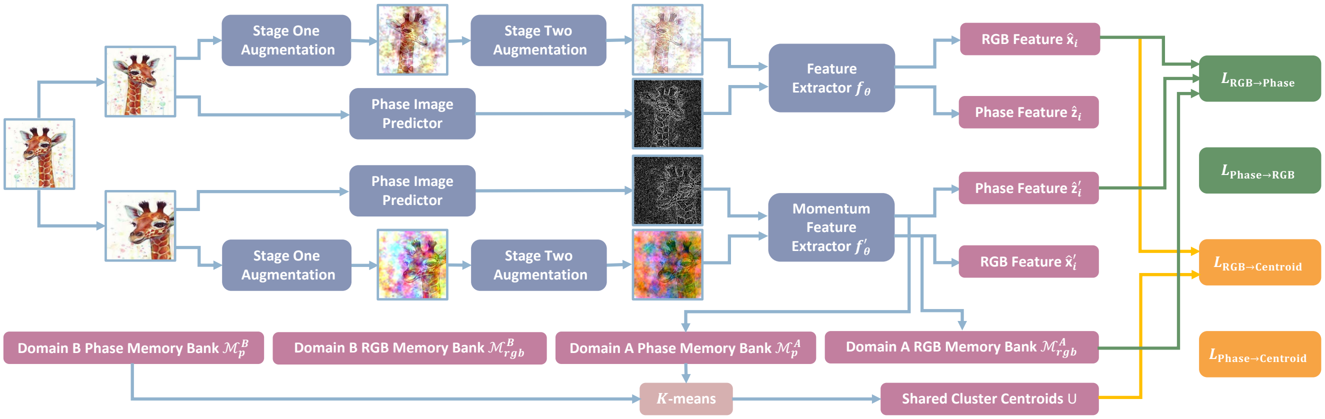
where $\mathrm{Im}(\cdot)$ and $\mathrm{Re}(\cdot)$ represent imaginary and real part of the input.

Stage One: Low-frequency Phase Augmentation. The first step of our proposed domain augmentation technique is perturbing the low-frequency components of phase PH_i while keeping the high-frequency portion unchanged. In this case, visually different but semantically same variants of the original image I_i are created and involved in training to prevent model from overfitting to training domains. To augment the PH_i, we randomly select another image I_j from the training set and mix the low-frequency part of PH_i and PH_j. The new phase $\widehat{\mathrm{PH}}_i$ is generated by:

$$\begin{aligned} \widehat{\mathrm{PH}}_i &= \mathrm{HPH}_i + \alpha\,\mathrm{LPH}_i + (1-\alpha)\,\mathrm{LPH}_j \\ &= (1-\mathrm{R})\circ\mathrm{PH}_i + \alpha\,\mathrm{R}\circ\mathrm{PH}_i + (1-\alpha)\,\mathrm{R}\circ\mathrm{PH}_j, \end{aligned}$$

where

$$\mathrm{R} = \begin{cases} 1, (u,v) \in [c_H - r : c_H + r, c_W - r : c_W + r] \\ 0, \quad\quad \text{others} \end{cases}. \tag{3}$$



Stage One Augmentation
Stage Two Augmentation
Phase Image Predictor
Phase Image Predictor
Stage One Augmentation
Stage Two Augmentation
Feature Extractor f_θ
Momentum Feature Extractor f'_θ
RGB Feature $\hat{\mathbf{x}}_i$
Phase Feature $\hat{\mathbf{z}}_i$
Phase Feature $\hat{\mathbf{z}}'_i$
RGB Feature $\hat{\mathbf{x}}'_i$
$L_{\text{RGB}\rightarrow\text{Phase}}$
$L_{\text{Phase}\rightarrow\text{RGB}}$
$L_{\text{RGB}\rightarrow\text{Centroid}}$
$L_{\text{Phase}\rightarrow\text{Centroid}}$
Domain B Phase Memory Bank $\mathcal{M}_p^B$
Domain B RGB Memory Bank $\mathcal{M}_{rgb}^B$
Domain A Phase Memory Bank $\mathcal{M}_p^A$
Domain A RGB Memory Bank $\mathcal{M}_{rgb}^A$
K-means
Shared Cluster Centroids U

Figure 3. Illustration of our framework. The shared cluster centroids $\mathbf{U}$ are calculated by applying K-means to all phase features in training domain memory banks. Only the inputs for $\mathcal{L}_{\text{RGB}\rightarrow\text{Phase}}$ and $\mathcal{L}_{\text{RGB}\rightarrow\text{Centroid}}$ are indicated in the figure for brevity. $\mathcal{L}_{\text{Phase}\rightarrow\text{RGB}}$ and $\mathcal{L}_{\text{Phase}\rightarrow\text{Centroid}}$ can be calculated in the same way by exchanging the role of RGB and Phase. Best viewed zoom-in and in color.

R is a low-pass filter. (c_H, c_W) and r indicate the image center and the boundary for low-frequency components. $\circ$ denotes the Hadamard product. HPH_i and LPH_i represent high-pass filtered and low-pass filtered phases for I_i. $\alpha \sim U(0, \lambda)$ controls the degree of augmentation.

Stage Two: Amplitude Augmentation. Existing research works [17, 23] have certified that the amplitude spectrum mainly contains the low-level statistics, and distorting the amplitude components would not affect the high-level semantics. Therefore, to further generate more diversified data for training, we augment the amplitude spectrum in the second step. Specifically, the amplitude perturbation is formulated as:

$$\widehat{\mathrm{AM}}_i = \beta\,\mathrm{AM}_i + (1-\beta)\,\mathrm{AM}_j, \tag{4}$$

where AM_j is the amplitude spectrum of the image I_j selected in the low-frequency augmentation step. $\beta \sim U(0, \eta)$ defines the strength of amplitude augmentation. The final augmented image becomes:

$$\begin{aligned}\hat{I}_i &= \mathcal{F}^{-1}(\hat{F}_i(u,v)) \\ &= \mathcal{F}^{-1}(\widehat{\mathrm{AM}}_i(u,v)e^{-j\widehat{\mathrm{PH}}_i(u,v)}).\end{aligned} \tag{5}$$

Here, $\mathcal{F}^{-1}(\cdot)$ is the inversed Fourier transformation. $\hat{F}_i(u,v)$ denotes the augmented Fourier representation.

3.2. Phase-enhanced Contrastive Learning

As shown in Fig. 2, the discrepancy between phase images from two domains is clearly smaller than the corresponding RGB images. Consequently, we propose to use phase image as a proxy to align features across domains, and enhance the vanilla contrastive learning to learn domain-agnostic and class-aware features for cross-domain image retrieval. The phase image PI_i is produced by phase image predictor with constant amplitude γ:

$$\mathrm{PI}_i = \mathcal{F}^{-1}(\gamma e^{-j\,\mathrm{PH}_i(u,v)}). \tag{6}$$

Instance-Instance Contrastive Learning. The vanilla contrastive learning methods are designed to extract semantic information by pulling different variants of the same RGB images closer while pushing the others (negatives) away, *i.e.*:

$$\begin{aligned}\mathcal{L}_{\text{rgb}} &= \mathrm{Contra}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}'_i, \hat{\mathbf{x}}'_e) \\ &= \sum_{i=0}^{M-1} -\log \frac{\exp(\hat{\mathbf{x}}_i^\top \hat{\mathbf{x}}'_i/\tau)}{\exp(\hat{\mathbf{x}}_i^\top \hat{\mathbf{x}}'_i/\tau) + \sum\limits_{e=0}^{E-1} \exp(\hat{\mathbf{x}}_i^\top \hat{\mathbf{x}}'_e/\tau)},\end{aligned} \tag{7}$$

where $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}'_i$ are the feature representations for two augmented views of image I_i. The negative image is denoted by $\hat{\mathbf{x}}'_e$, and E is the number of selected negatives stored in the feature memory bank, following the same strategy as detailed in [3]. τ is a temperature hyper-parameter. However, the vanilla contrastive learning loss overlooks the effect of domain-specific knowledge, and as illustrated in [33], the contrastive training is likely to collapse in those cases with large domain shifts. We thus introduce phase images to bridge the domain gap and encourage domain-invariant feature learning. To effectively assist category-level cross-domain alignment, it is essential to guarantee that the phase features are meaningful. We achieve this by applying contrastive loss to the phase images:

$$\mathcal{L}_{\text{ph-intra}} = \mathrm{Contra}(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}'_i, \hat{\mathbf{z}}'_e) \tag{8}$$

In $\mathcal{L}_{\text{ph-intra}}$, the two augmented RGB images are replaced with their phase images PI_i and PI'_i, and the same feature extractor f_θ and momentum feature extractor f'_θ are employed to predict the corresponding features $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}'_i$ for the phase images. Owing to the small domain discrepancy between phase images, we can then mitigate the gap between original RGB images from different domains by pulling the RGB features and paired phase features closer with RGB-Phase contrastive loss $\mathcal{L}_{\text{ph-cross}}$:

$$\begin{aligned}\mathcal{L}_{\text{ph-cross}} &= \frac{1}{2}(\mathcal{L}_{\text{RGB}\to\text{Phase}} + \mathcal{L}_{\text{Phase}\to\text{RGB}}) \\ &= \frac{1}{2}(\text{Contra}(\hat{\mathbf{x}}_i, \hat{\mathbf{z}}'_i, \hat{\mathbf{x}}'_e) + \text{Contra}(\hat{\mathbf{z}}_i, \hat{\mathbf{x}}'_i, \hat{\mathbf{z}}'_e))\end{aligned} \tag{9}$$

The objective of $\mathcal{L}_{\text{RGB}\to\text{Phase}}$ is to make sure the query RGB feature $\hat{\mathbf{x}}_i$ is closer to the phase feature $\hat{\mathbf{z}}'_i$ compared to the distance to other negative RGB feature $\hat{\mathbf{x}}'_e$ from RGB memory bank. A feature memory bank is maintained separately for each domain. In $\mathcal{L}_{\text{Phase}\to\text{RGB}}$, we use phase feature $\hat{\mathbf{z}}_i$ as query instead. $\mathcal{L}_{\text{ph-intra}}$ and $\mathcal{L}_{\text{ph-cross}}$ then corporate together to facilitate domain-invariant feature learning:

$$\mathcal{L}_{\text{ph-i}} = \mathcal{L}_{\text{ph-intra}} + \mathcal{L}_{\text{ph-cross}}. \tag{10}$$

Instance-Centroid Contrastive Learning. Both $\mathcal{L}_{\text{rgb}}$ and $\mathcal{L}_{\text{ph-i}}$ focus on the instance-wise contrastive learning where features for image I_i are separated from all the other instances. Nevertheless, image features of the same class should be clustered together in our category-level cross-domain image retrieval task, which inspires us to propose phase-enhanced instance-centroid contrastive loss. Here, we introduce a set of cluster centroids calculated based on only phase features to enable cross-domain sharing. Different from domain-dependent RGB features, phase features with smaller domain discrepancies are intuitively more suitable for centroids measurement. With the shared cluster centroids, we can then pull samples with similar semantic information closer and encourage the shared semantic structure encoding across domains. More concretely, we apply K-means to all features in both phase memory bank, *i.e.* $\mathcal{M}_p^A$ and $\mathcal{M}_p^B$ to predict K cluster centroids $\mathbf{U} = \{\mathbf{u}_k\}_{k=1}^{K}$. The corresponding cluster centroid for image I_i is denoted as $\mathbf{u}_i$, and the instance-centroid contrastive loss is:

$$\begin{aligned}\mathcal{L}_{\text{ph-c}} &= \frac{1}{2}(\mathcal{L}_{\text{RGB-Centroid}} + \mathcal{L}_{\text{Phase-Centroid}}) \\ &= \frac{1}{2}\sum_{i=0}^{M-1} -(\log \frac{\exp(\hat{\mathbf{x}}_i^\top \hat{\mathbf{u}}_i/\phi)}{\sum_{k=0}^{K-1}\exp(\hat{\mathbf{x}}_i^\top \hat{\mathbf{u}}_k/\phi)} + \\ &\qquad \log \frac{\exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{u}}_i/\phi)}{\sum_{k=0}^{K-1}\exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{u}}_k/\phi)}),\end{aligned} \tag{11}$$

where ϕ is a temperature hyper-parameter. We apply all the training losses to both training domains.

4. Experiments

4.1. Datasets and Settings

Datasets. To evaluate the efficacy of our proposed method, extensive experiments are carried out using PACS [13] and DomainNet [21] dataset. PACS consists of four different domains (Photo, Art Painting, Cartoon, Sketch) with seven categories. DomainNet offers six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. We use the same categories selected in [9] for evaluation.

Implementation Details. We adopt ResNet-50 [8] as the feature extractor θ, which is initialized with the parameters of pre-trained unsupervised MoCov2 [3] model to guarantee no labeled data is used in the whole training process. The learning rate for feature extractor is set to 0.0002 at the beginning and then gradually decreases according to the cosine learning rate schedule. The size of RGB and phase memory bank for each domain is 5120. Batch size of 64 is fixed in all experiments for fair comparisons. The number of centroids K is assigned according to the number of training categories. The boundary r for low-frequency components is set to 25, and we maintain a constant amplitude γ at 5e4. Our framework is trained using SGD optimizer with 0.9 momentum factor and 1e-4 weight decay, and implemented based on the deep learning library PyTorch [19].

Evaluation Metrics. We use the same retrieval precision (P@50, P@100 and P@200) as [9] to validate the performance of our framework. Notably, we utilize the trained feature extractor to extract features from test domain RGB images, which serve as the sole input for testing. The retrieval is performed by calculating the cosine distance between the features of RGB images. Taking cartoon image retrieval with a query sketch guitar as an example, we first calculate the cosine distances between the sketch guitar feature and all image features from cartoon domain, and then rank all cartoon images according to the cosine distance in ascending order. P@50 measures the precision of retrieving cartoon guitars among the top 50 retrievals since only those from the same category as the query are correct retrievals. P@100 and P@200 provide a more thorough evaluation as the top 100 and 200 are considered.

Baselines. To analyze the effectiveness of our method, we use the following baselines: 1) CDS [12] is a self-supervised learning method designed for muti-domain data. Semantic-aware and domain-invariant feature learning is achieved by the proposed in-domain instance discrimination and cross-domain matching method. 2) PCS [33] introduces prototypes to assist semantic feature encoding and alignment across the domains in an unsupervised manner; 3) UCDIR [9] targets at unsupervised cross-domain image retrieval. A distance-of-distance loss is proposed to

Table 1. Cross-domain retrieval accuracy (%) on PACS dataset.

Train	Art paint., Cartoon			Art paint., Photo			Art paint., Sketch			Cartoon, Photo		
Test	Photo, Sketch			Cartoon, Sketch			Cartoon, Photo			Art paint., Sketch		
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
CDS [12]	20.06	18.45	16.50	27.20	25.67	24.78	41.84	38.19	34.12	19.64	19.64	19.48
PCS [33]	36.47	34.16	31.59	26.74	25.19	24.58	39.55	35.54	31.50	26.19	24.44	22.27
UCDIR [9]	28.13	25.48	23.22	28.13	26.45	25.52	41.02	37.15	31.96	28.78	26.46	24.14
USBIR [10]	44.16	42.34	40.38	35.08	32.97	31.09	31.74	30.31	28.46	37.74	34.78	31.54
BrAD [7]	42.84	40.02	37.16	36.54	33.75	31.75	52.91	49.59	44.48	35.83	33.72	31.53
Ours	**56.88**	**53.83**	**50.28**	**43.87**	**41.02**	**38.30**	**58.17**	**54.82**	**49.38**	**51.70**	**49.48**	**46.60**

Train	Cartoon, Sketch			Photo, Sketch			Average			Improvement		
Test	Art paint., Photo			Art paint., Cartoon								
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
CDS [12]	50.14	44.94	38.66	35.22	33.10	33.10	32.35	29.97	27.35	+23.07	+22.34	+20.38
PCS [33]	56.50	50.89	42.65	32.78	29.49	26.11	36.37	33.29	29.78	+19.05	+19.02	+17.95
UCDIR [9]	56.38	50.62	42.30	34.09	30.62	27.38	36.09	32.80	29.09	+19.33	+19.51	+18.64
USBIR [10]	47.56	45.16	42.64	30.14	28.43	26.61	37.74	35.67	33.45	+17.68	+16.64	+14.28
BrAD [7]	59.01	54.16	46.54	44.87	41.85	37.47	45.33	42.18	38.16	+10.09	+10.13	+9.57
Ours	**68.85**	**65.24**	**57.31**	**53.02**	**49.49**	**44.53**	**55.42**	**52.31**	**47.73**	/	/	/

Table 2. Cross-domain retrieval accuracy (%) on DomainNet dataset.

Train	Clipart, Sketch			Info., Real			Info., Sketch			Paint., Clipart		
Test	Info., Real, Paint., Quick.			Clipart, Sketch, Paint., Quick.			Clipart, Real, Paint., Quick.			Info., Sketch, Real, Quick.		
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
CDS [12]	37.62	35.18	32.10	35.90	32.57	27.85	44.31	41.08	36.70	35.18	32.18	28.73
PCS [33]	38.08	35.24	32.21	40.43	36.81	32.27	41.24	37.98	34.20	36.34	33.93	30.82
UCDIR [9]	40.52	37.21	33.95	41.53	38.42	34.20	41.74	39.09	35.44	39.16	36.90	33.59
USBIR [10]	46.35	44.60	41.71	48.17	46.15	42.51	51.54	49.77	46.92	50.76	48.21	44.69
BrAD [7]	44.62	41.61	38.01	49.99	46.72	41.57	52.25	49.73	45.67	43.54	40.73	36.80
Ours	**52.91**	**50.59**	**47.09**	**63.48**	**62.10**	**58.86**	**61.14**	**59.17**	**55.52**	**53.29**	**51.14**	**47.31**

Train	Paint., Quick.			Quick., Real			Average			Improvement		
Test	Clipart, Sketch, Info., Real			Paint., Clipart, Info., Sketch								
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
CDS [12]	47.71	41.70	34.79	38.97	34.04	28.68	39.95	36.13	31.48	+19.74	+20.97	+21.11
PCS [33]	49.08	43.69	37.40	41.39	36.50	31.04	41.09	37.36	32.99	+18.60	+19.74	+19.60
UCDIR [9]	53.00	48.56	42.80	44.99	40.15	34.63	43.49	40.06	35.77	+16.20	+17.04	+16.82
USBIR [10]	38.93	35.00	30.18	56.84	53.08	47.16	48.77	46.14	42.20	+10.92	+10.96	+10.39
BrAD [7]	59.17	54.21	46.85	55.90	50.91	43.68	50.91	47.32	42.10	+8.78	+9.78	+10.49
Ours	**64.16**	**59.97**	**53.30**	**63.18**	**59.60**	**53.44**	**59.69**	**57.10**	**52.59**	/	/	/

mitigate the domain discrepancy. 4) USBIR [10] is originally designed for unsupervised sketch-based image retrieval, where cross-domain correspondence is estimated by the prototype and feature memory bank-enhanced optimal transport. 5) BrAD [7] works towards the unsupervised domain generalization. An auxiliary domain is devised to remedy the domain shifts and learn domain-agnostic features.

4.2. Main Results

4.2.1 PACS Dataset

Settings. Any two out of the four domains (Photo, Art painting, Cartoon, and Sketch) in the PACS dataset are selected as one training pair. There are six training pairs in total, and the remaining two domains are employed as the unseen test domains. For instance, the test domains are Photo and Sketch when the model is trained with data from Art painting and Cartoon. We then conduct both photo → sketch and sketch → photo retrieval and calculate the mean of the precisions as the final results.

Results. The retrieval results in Table 1 demonstrate that: 1) Existing multi-domain unsupervised feature learning methods [12, 33] cannot generalize well to novel domains. 2) Unsupervised cross-domain image retrieval algorithms [9, 10] designed for a specific domain pair also suffer from the large domain shifts between seen and unseen domains. 3) BrAD [7] performs well in DG-UCDIR task among the existing baselines. 4) Our framework achieves the best results in all six pairs in terms of P@50, P@100, and P@200, and the overall average scores improve the others by a large margin. 5) Generalizing from Art painting and Photo to Cartoon and Sketch is the most difficult setting, resulting in 43.87% at P@50 for our method.

4.2.2 DomainNet Dataset

Settings. Following the experiment settings in [9], we use the six different pairs for model training. While training with one domain pair, the rest four domains in the Do-

Table 3. Contribution of each proposed component to cross-domain retrieval accuracy $(\%)$ on DomainNet dataset.

Train	Clipart, Sketch			Info., Real			Info., Sketch			Paint., Clipart		
Test	Info., Real, Paint., Quick.			Clipart, Sketch, Paint., Quick.			Clipart, Real, Paint., Quick.			Info., Sketch, Real, Quick.		
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
L_{rgb} (v1)	40.85	38.00	34.57	46.05	42.61	37.63	47.54	44.69	40.68	38.37	35.58	31.90
v1+Stage one Aug. (v2)	43.33	40.33	36.71	50.34	47.30	42.33	51.22	48.76	44.54	41.65	38.76	34.97
v2+Stage two Aug. (v3)	47.04	43.94	40.12	54.54	51.81	46.97	56.75	54.30	49.89	44.26	41.29	37.19
v3+L_{ph-i} (v4)	49.92	47.02	43.15	59.11	56.39	51.42	59.81	57.60	53.45	50.04	47.21	43.04
v4+L_{ph-c} (v5)	**52.91**	**50.59**	**47.09**	**63.48**	**62.10**	**58.86**	**61.14**	**59.17**	**55.52**	**53.29**	**51.14**	**47.31**

Train	Paint., Quick.			Quick., Real			Average			Component Contribution		
Test	Clipart, Sketch, Info., Real			Paint., Clipart, Info., Sketch								
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
L_{rgb} (v1)	52.38	47.23	40.52	50.33	44.92	38.06	45.92	42.17	37.23	/	/	/
v1+Stage one Aug. (v2)	58.45	53.63	46.32	55.93	50.97	43.80	50.15	46.63	41.45	+4.23	+4.46	+4.22
v2+Stage two Aug. (v3)	61.01	56.41	49.06	59.03	54.35	47.16	53.77	50.35	45.07	+3.62	+3.72	+3.62
v3+L_{ph-i} (v4)	63.77	59.54	52.73	62.04	57.78	50.87	57.45	54.26	49.11	+3.68	+3.91	+4.04
v4+L_{ph-c} (v5)	**64.16**	**59.97**	**53.30**	**63.18**	**59.60**	**53.44**	**59.69**	**57.10**	**52.59**	+2.24	+2.84	+3.48

Table 4. Influence of different seen domains over cross-domain retrieval accuracy $(\%)$ on DomainNet dataset.

Test	Clipart, Sketch			Test	Paint., Clipart			Test	Paint., Quick.		
Train	P@50	P@100	P@200	Train	P@50	P@100	P@200	Train	P@50	P@100	P@200
Quick., Real	73.91	69.53	62.23	Quick., Real	74.12	69.96	63.73	Info., Sketch	41.54	40.17	38.02
Paint., Quick.	74.92	70.72	62.28	Info., Sketch	75.09	72.16	65.63	Clipart, Sketch	43.71	43.33	42.04
Info., Real	**76.91**	**74.23**	**68.49**	**Info., Real**	**76.29**	**74.35**	**69.37**	**Info., Real**	**47.61**	**47.39**	**46.31**

Test	Info., Real			Test	Info., Sketch			Test	Quick., Real		
Train	P@50	P@100	P@200	Train	P@50	P@100	P@200	Train	P@50	P@100	P@200
Paint., Quick.	53.27	47.74	41.50	Quick., Real	49.42	45.83	39.89	Info., Sketch	43.01	41.59	39.87
Clipart, Sketch	55.67	50.00	42.99	Paint., Quick.,	52.21	46.36	39.20	Clipart, Sketch	44.43	44.18	43.17
Paint., Clipart	**57.99**	**52.93**	**44.95**	**Paint., Clipart**	**56.87**	**52.11**	**43.69**	**Paint., Clipart**	**45.59**	**45.10**	**43.99**

mainNet dataset [21] are all regarded as unseen test domains. For example, the retrieval evaluations are carried out among Infograph, Real, Painting and Quickdraw if Clipart and Sketch are the seen domains, and the retrieval results are the averaged precisions of 12 retrieval tasks (Infograph → Real, Real → Infograph, Infograph → Painting, Painting → Infograph, Infograph → Quickdraw, Quickdraw → Infograph, Real → Painting, Painting → Real, Real → Quickdraw, Quickdraw → Real, Painting → Quickdraw and Quickdraw → Painting).

Results. From the retrieval results in Table 2, we draw the following conclusions: 1) The domain gaps between seen and unseen domains in the DomainNet dataset also hurts the retrieval accuracies of the baselines, but BrAD [7] still performs the best among them. 2) Our framework significantly outperforms all the compared methods in all six settings. 3) When taking Painting and Quickdraw for training, the retrieval accuracies for the remaining test domains are the highest, *i.e*. 64.16% for P@50. 4) Transferring from Infograph and Real domain gives us the most significant improvement over the best baseline, and the retrieval accuracy increases from 49.99% to 63.48% in terms of P@50. Qualitative retrieval results can be found in the supplementary.

4.3. Ablation Study

The contribution of each proposed component is comprehensively evaluated on all six DomainNet experimental settings. The results in Table 3 show: 1) Applying vanilla contrastive learning on RGB images (v1) overlooks the influence of large domain gap and thus results in unsatisfactory performance. 2) Augmentation in low-frequency phase part (v2) improves the novel domain generalizability of the model. 3) In v3, the second augmentation step (amplitude augmentation) boosts the retrieval performance by 3.62% for P@50. 4) With phase-enhanced instance-instance contrastive loss (v4), features from different domains are better aligned and the performance gain over v3 is 3.68% for P@50. 5) Our full model (v5), which further takes the advantage of phase-enhanced instance-centroid contrastive loss, predicts the best feature for DG-UCDIR.

4.4. Influence of Different Seen Domains

To analyze the effect of training domains, we measure the retrieval accuracy for the same test domains by applying models trained with different seen domain pairs. For instance, the reported retrieval precisions for Clipart-Sketch test pair in Table 4 are calculated based on three models learned with Quickdraw-Real, Painting-Quickdraw, and Infograph-Real respectively. From Table 4 we can see: 1) The retrieval performance for the same test domains varies regarding the training domains. For Infograph-Sketch retrieval, the model trained with Painting and Clipart achieves 56.87% at P@50, while its Quickdraw-Real counterpart is only 49.42%. 2) The retrieval accuracies for different un-

Table 5. Cross-category DG-UCDIR accuracy (%) on DomainNet dataset.

Train	D-Clipart, D-Sketch			D-Info., D-Real			D-Info., D-Sketch			D-Paint., D-Clipart		
Test	P-Photo, P-Sketch, P-Art paint.			P-Cartoon, P-Sketch, P-Art paint.			P-Cartoon, P-Sketch, P-Art paint., P-Photo			P-Photo, P-Sketch		
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
CDS [12]	33.30	30.52	27.09	21.68	20.20	19.44	26.92	25.08	23.20	18.42	16.50	14.87
PCS [33]	29.97	27.04	23.96	25.73	24.09	22.99	29.56	27.13	24.75	21.36	18.74	15.97
UCDIR [9]	29.57	26.38	23.62	25.83	24.12	23.10	27.35	25.50	23.91	23.95	20.63	18.62
USBIR [10]	29.73	27.76	26.33	25.78	24.69	24.20	29.20	27.77	26.44	18.54	17.04	16.86
BrAD [7]	37.35	34.24	30.83	31.97	29.79	27.93	38.77	36.17	33.15	30.30	27.68	25.11
Ours	**39.73**	**36.00**	**32.45**	**34.59**	**32.02**	**29.77**	**41.57**	**38.22**	**34.98**	**34.79**	**30.89**	**28.03**
Train	D-Paint., D-Quick.			D-Quick., D-Real			Average			Improvement		
Test	P-Photo, P-Cartoon			P-Art paint., P-Cartoon								
Metrics	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
CDS [12]	34.18	32.71	30.34	27.42	25.18	23.14	26.99	25.03	23.01	+11.52	+10.31	+9.36
PCS [33]	30.38	28.80	27.29	25.97	23.20	20.94	27.16	24.83	22.65	+11.35	+10.51	+9.72
UCDIR [9]	34.02	32.21	30.73	29.62	26.63	23.62	28.39	25.91	23.93	+10.12	+9.43	+8.44
USBIR [10]	28.99	26.98	24.52	29.50	27.84	26.12	26.96	25.35	24.08	+11.55	+9.99	+8.29
BrAD [7]	40.63	38.82	36.09	33.90	31.35	28.62	35.49	33.01	30.29	+3.02	+2.33	+2.08
Ours	**41.97**	**39.02**	**36.17**	**38.42**	**35.91**	**32.81**	**38.51**	**35.34**	**32.37**	/	/	/

seen pairs differ greatly even when tested with the same model. When evaluated with the model trained on Painting and Clipart, P@50 for Quickdraw-Real test pair with a larger domain gap is 12.40% lower than Infograph-Real.

4.5. Results of Cross-category DG-UCDIR

Settings. Cross-category DG-UCDIR is an even harder setting with test images of unseen categories from novel domains. To verify whether our framework can also generalize to novel categories, we apply the model trained on DomainNet dataset to conduct retrieval between PACS domains since there is no class overlap between the DomainNet subset we used [9] and the 7 categories in PACS dataset. Furthermore, as we can see from Fig. 4, the four domains in the PACS dataset can all find their visually similar domains from the DomainNet dataset. More examples are provided in the supplementary file. The four pairs are: P-Sketch and D-Quickdraw, P-Cartoon and D-Clipart, P-Art painting and D-Painting, P-Photo and D-Real. P- and D- represent PACS and DomainNet dataset, respectively. To simulate the cross-category DG-UCDIR scenario, the visually similar domains are excluded during test time. Taking the model trained with D-Quickdraw and D-Real as an example, the test domains are just the remaining two domains (P-Art painting and P-Cartoon) from PACS dataset since D-Quickdraw and D-Real are paired with P-Sketch and P-Photo.

Results. From the results shown in Table 5, we can draw the conclusions: 1) Cross-category DG-UCDIR is a more challenging task than DG-UCDIR. P@50 for retrieval between P-Photo and P-Sketch is 34.79% by adopting our model trained with D-Painting and D-Clipart as a feature extractor. However, the corresponding retrieval accuracy is 56.88% (*c.f.* Table 1) under the circumstance without category generalization when we replace D-Painting and D-Clipart with their paired domains (P-Art painting and P-Cartoon) from PACS dataset. 2) Our framework still achieves the best averaged retrieval accuracies compared with all baselines, which shows the superior performance of our method in transferring knowledge from seen classes to unseen classes. 3) It is worthwhile to explore further in the cross-category DG-UCDIR task and equip model with better category generalization capability for real applications.

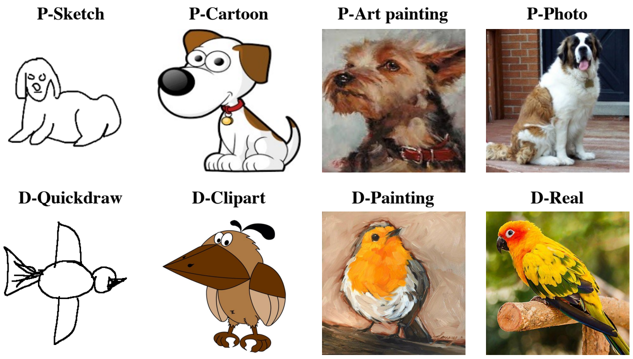


Figure 4. Examples from PACS and DomainNet dataset.

5. Conclusion

In this paper, we propose a new research problem of domain-generalized unsupervised cross-domain image retrieval (DG-UCDIR) under the stringent assumption of unseen test domain and no annotated data. We believe that this new research direction would provide a step towards more practical cross-domain image retrieval applications. To facilitate cross-domain test, we introduce a novel two-stage domain augmentation strategy to enrich the diversity of training data. The phase-enhanced instance-instance and instance-centroid contrastive losses are proposed for domain-agnostic and semantic-aware feature learning. Extensive experiments are conducted on the PACS and DomainNet datasets to provide insights into factors that affect DG-UCDIR performance.

Acknowledgements

This research 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] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv*, 2020.

[4] Ilke Cugu, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *CVPR*, 2022.

[5] Thomas Duboudin, Emmanuel Dellandréa, Corentin Abgrall, Gilles Hénaff, and Liming Chen. Encouraging intraclass diversity through a reverse contrastive loss for single-source domain generalization. In *ICCV*, 2021.

[6] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *TOG*, 2012.

[7] Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, Dina Katabi, Kate Saenko, Rogerio S. Feris, and Leonid Karlinsky. Unsupervised domain generalization by learning a bridge across domains. In *CVPR*, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *ECCV*, 2022.

[10] Conghui Hu, Yongxin Yang, Yunpeng Li, Timothy M Hospedales, and Yi-Zhe Song. Towards unsupervised sketch-based image retrieval. In *BMVC*, 2022.

[11] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015.

[12] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *ICCV*, 2021.

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[14] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *ICLR*, 2020.

[15] Yadan Luo, Ziwei Wang, Zi Huang, Yang Yang, and Huimin Lu. Snap and find: deep discrete cross-domain garment image retrieval. *arXiv preprint arXiv:1904.02887*, 2019.

[16] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, 2022.

[17] A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. Phase in speech and pictures. In *ICASSP*, 1979.

[18] Shuxin Ouyang, Timothy M. Hospedales, Yi-Zhe Song, and Xueming Li. Forgetmenot: Memory-aware forensic facial sketch matching. In *CVPR*, 2016.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

[20] Chunlei Peng, Xinbo Gao, Nannan Wang, Dacheng Tao, Xuelong Li, and Jie Li. Multiple representations-based face sketch–photo synthesis. *TNNLS*, 2016.

[21] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[22] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *NeurIPS*, 2016.

[23] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 1982.

[24] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021.

[25] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *TOG*, 2016.

[26] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

[27] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.

[28] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *NeurIPS*, 2020.

[29] Jingye Wang, Ruoyi Du, Dongliang Chang, Kongming Liang, and Zhanyu Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *ACM MM*, 2022.

[30] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, 2021.

[31] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[32] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Fine-grained instance-level sketch-based image retrieval. *IJCV*, 2021.

[33] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *CVPR*, 2021.