

Universal Cross-Domain Retrieval: Generalizing Across Classes and Domains

Soumava Paul^{1†}, Titir Dutta^{2*}, Soma Biswas²

¹Indian Institute of Technology, Kharagpur, ²Indian Institute of Science, Bangalore

soumava2016@gmail.com, f titird,somabiswas g@iisc.ac.in

Abstract

In this work, for the first time, we address the problem of universal cross-domain retrieval, where the test data can belong to classes or domains which are unseen during training. Due to dynamically increasing number of categories and practical constraint of training on every possible domain, which requires large amounts of data, generalizing to both unseen classes and domains is important. Towards that goal, we propose SnMpNet (Semantic Neighborhood and Mixture Prediction Network), which incorporates two novel losses to account for the unseen classes and domains encountered during testing. Specifically, we introduce a novel Semantic Neighborhood loss to bridge the knowledge gap between seen and unseen classes and ensure that the latent space embedding of the unseen classes is semantically meaningful with respect to its neighboring classes. We also introduce a mix-up based supervision at image-level as well as semantic-level of the data for training with the Mixture Prediction loss, which helps in efficient retrieval when the query belongs to an unseen domain. These losses are incorporated on the SE-ResNet50 backbone to obtain SnMpNet. Extensive experiments on two large-scale datasets, Sketchy Extended and DomainNet, and thorough comparisons with state-of-the-art justify the effectiveness of the proposed model.

1. Introduction

Due to the availability of large amount of data in different domains of multi-media, cross-domain retrieval has gained significant attention. It addresses the challenging problem of retrieving relevant data from a domain (say, image), when the query belongs to a different domain (e.g. sketch, painting etc.). As motivation for our work, we focus on the specific application of sketch-based image retrieval (SBIR) [14][34], which has wide range of applications in e-commerce, forensic data matching, etc. Considering the dynamic real-world,

where the search dataset is always being augmented with new categories of data, recently, the focus has shifted to zero-shot SBIR (ZS-SBIR) or generalized ZS-SBIR (GZS-SBIR) [26][32][4][15][5][7], in which, the query and search set samples belong to classes not seen during training.

The generic architecture for ZS-SBIR (or other cross-domain retrieval applications) consists of two parallel branches, each consisting of a feature extractor and a classifier, to learn the latent-space representations of the data from individual domains (here, sketches and images). The domain gap in this latent space is bridged by the semantic descriptions [19][22] of the seen classes. During testing, the query sketch and search set images are projected to this space and compared directly for retrieval. But if the query belongs to a different domain, say painting, then this network needs to be re-trained, with painting and images as the two domains. This not only requires the training to be performed for every domain-pair with sufficient amount of data, but also, the query domain needs to be known a-priori.

Here, we attempt a more realistic and significantly more challenging cross-domain retrieval scenario, where the query data can belong not only to unseen classes, but also to unseen domains - which we term as universal cross-domain retrieval (UCDR). It is a combination of two well-studied, but separate problems in literature, namely, ZS-SBIR, which accounts for test data from unseen classes and domain generalization (DG) [28][10], which accounts for test data from unseen domains in the classification. To this end, we propose SnMpNet (Semantic Neighbourhood and Mixture Prediction Network) which is a single-branch network consisting of a feature extractor and classifier, for learning a domain-independent embedding of input data, that also generalizes to unseen category test data. For generalizing to unseen classes, we propose Semantic Neighbourhood Loss to represent the unseen classes in terms of their relative positions with the seen classes. In addition, we exploit the mix-up technique [33] to populate our training set with samples created through both inter-class and inter-domain mixing to prepare for unseen domains of query samples during testing. To better generalize across domains, we propose a novel Mixture Prediction Loss. The contributions

* Equal contribution.

of this work are summarized below:

- (1) We propose a novel framework SnMpNet to address the universal cross-domain retrieval scenario, where the query data may belong to seen/unseen classes, along with seen/unseen domains. To the best of our knowledge, this is the first work in literature addressing this extremely challenging problem.
- (2) We propose two novel losses, namely Semantic Neighbourhood loss and Mixture Prediction loss, to account for unseen classes and unseen domains during testing.
- (3) Extensive experiments and analysis of the proposed framework on Sketchy-Extended [24] and DomainNet [21] datasets are reported along with other state-of-the-art approaches modified for this application.

2. Related Work

We first describe the recent advancements in ZS-SBIR and DG, since UCDR can be considered as a combination of these. We also discuss the recently proposed classification-protocol for samples from unseen-class and domains and explain its differences with UCDR.

Zero-shot Sketch-based Image Retrieval (ZS-SBIR): ZS-SBIR protocol was first proposed in [26][32]. Later, several algorithms [5][7][4][6][31] have been proposed to address ZS-SBIR and its generalized version GZS-SBIR. All these algorithms follow the standard architecture with two parallel branches. In contrast, [15] proposes a single branch of network for processing data from both domains, along with a domain-indicator to embed the domain-discriminating information. All these algorithms use semantic-information [19][22] to account for the knowledge-gap, an idea which is inspired from zero-shot learning (ZSL) [29], which we will discuss later. UCDR generalizes the task of ZS-SBIR to additionally handle unseen domains during retrieval.

Domain Generalization (DG): DG refers to the task of classifying data from unseen domains, when the network has been trained with data from several other domains belonging to the same classes. This is usually addressed by learning a domain-invariant feature representation of the data, using techniques like self-supervision [2], triplet loss [28], maximum mean discrepancy (MMD) [20] loss, adversarial loss [11]. Recently, meta-learning [9] and episodic training [10] have shown impressive performance for the DG task. UCDR generalizes DG to additionally handle unseen classes in a retrieval framework.

Zero-shot Domain Generalization (ZSDG): Our work is also related to the well-researched Zero-shot Learning (ZSL) task, where the goal is to classify images from unseen classes during testing. Several seminal works have been proposed for this problem [23][1][35][25][30][29].

The knowledge gap between the seen and unseen classes is bridged using their corresponding semantic information.

Recently, few works have addressed the more realistic ZSDG task, which aims to classify unseen classes across generalized domains [17][18]. A mix-up based network is proposed in [17]. The work in [18] extends domain-generalization methods, e.g. feature-critic network [13], multi-task auto-encoder [8] to classify unseen-class samples by incorporating the semantic-information into their existing architecture. Recently, [27] discusses a retrieval protocol from any source domain to any target domain, using dedicated convnets for each of the training domains. UCDR extends the ZSDG-protocol to a retrieval framework. In contrast to ZSL or ZSDG protocol, no semantic information of the unseen classes are exploited in UCDR. This makes the UCDR protocol even more realistic and challenging, since in real-world, we may not have apriori information as to which classes will be encountered during testing.

3. Problem Definition

First, we define the task of universal cross-domain retrieval (UCDR), and the different notations used. We assume that labeled data from M -different ($M \geq 2$) domains (image, clip-art, painting, etc.) are available for training as, $D_{train} = \{x_i^{c,d}; c \in \{1, \dots, N_d\}, d \in \{1, \dots, D_g\}\}$. Here, $x_i^{c,d}$ is the i^{th} sample from d^{th} -domain, which belongs to c^{th} class. N_d is the number of examples in the d^{th} domain. Clearly, $M = 2$ represents the training set for standard cross-domain retrieval. The class labels for all the domains belong to unseen class set C_{train} . The goal is to find a latent domain-independent subspace, \mathbb{R}^m , such that, samples from the same class across all domains come closer and samples from different classes are pushed away in this space. Thus, for a query x_q and a search set $S_s = \{x_s; s=1, \dots, N_s\}$, we can retrieve the relevant search-set data using nearest neighbour of the query sample, projected in this learned subspace.

The proposed UCDR protocol is a combination of two separate experimental frameworks, namely: 1) U^d CDR - where the query x_q belongs to an unseen class, but seen domains $d \in \{1, \dots, D_g\}$. This implies that $C_{train} \setminus C_{test} = \{x_q\}$, where C_{test} is the set of possible classes of x_q , 2) U^c CDR - where the domain of x_q is unseen but the class is seen, i.e., $d \in \{1, \dots, D_g\}$, but $C_{test} = C_{train}$. The proposed combined protocol, where both the classes and domains can be unseen is denoted as U^d CDR, or simply UCDR (to avoid notational clutter). ZS-SBIR is a special case of U^d CDR, where sketch and real-images are the two domains. Also, U^d CDR is extension of DG protocol towards retrieval.

4. Proposed Approach

Here, we describe the proposed framework SnMpNet as details for addressing the UCDR task. SnMpNet is a single branch network consisting of a feature-extractor

and a classifier. Our main contributions are the semantic neighbourhood loss to account for unseen classes, and mixture prediction loss to account for unseen domains, integrated with a base network.

Proposed SnMpNet Framework - Overview: The proposed architecture for SnMpNet is illustrated in Figure 1. For this work, we choose SE-ResNet50 [16] as the backbone module for SnMpNet, motivated by its state-of-the-art performance for ZS-SBIR task [15]. Additionally, we incorporate an attention mechanism on top of this backbone, as in [4]. The embedding obtained for the input sample, $x_i^{c;d}$ from the base network is denoted as $g_i^{c;d} = \text{bb}(x_i^{c;d})$. This embedding is passed through the linear Mixture Prediction layer M_p , which ensures that $g_i^{c;d}$ is domain-invariant. Next, this domain-invariant feature is passed through the linear Semantic Neighbourhood layer, S_n to obtain the m -dimensional latent space representation $f_i^{c;d} = S_n(g_i^{c;d}) \in \mathbb{R}^m$. This m -dimensional space is the latent space, where we obtain semantically meaningful domain-independent representations of the data and effectively perform retrieval during testing.

The learning of this ϕ -space is driven by two objectives: 1) Unseen-class representation: We want to represent data from unseen classes (during testing) in a semantically meaningful manner in this space, taking into account the neighbourhood information. This is handled by the Semantic Neighbourhood loss (L_{S_n}). 2) Domain-independent representation: We want the ϕ -space representation to be independent of the domain of the input data, so that SnMpNet can accommodate data from unseen domains. This is addressed by the Mixture Prediction loss (L_{M_p}). We further incorporate an inter-class and inter-domain mix-up, to generate mixed samples and maintain only the categorical discrimination in ϕ -space by minimizing the Mixup Classification loss (L_{CE}^{mix}). Next, we describe the individual loss components to address the above objectives.

4.1. Unseen-class representation

The main challenge in handling unseen classes is to effectively and meaningfully embed them in the latent feature space ϕ , without any prior knowledge of those classes. Here, we propose to learn the ϕ -space embeddings of the training samples, so that they are semantically meaningful with respect to the other seen classes, especially its neighbour classes. Thus, during testing, the model learns to embed the unseen class query samples according to their semantic relevance into the ϕ -space. This is in contrast to the cross-entropy loss used for classification, or the standard metric learning losses, like triplet loss used for retrieval, where the goal is to bring data from the same class closer and move those from other classes far apart. Previous attempts to include the neighbourhood information in novel

the embedding space has been reported: Stochastic Neighbourhood Embedding [3], Memory-based Neighbourhood Embedding [12] etc. Here, we propose a novel Semantic Neighbourhood loss for this task as described next.

We learn the feature $f_i^{c;d}$, such that its distance with respect to the seen classes is same as the distance between its class-semantics and the semantics of the other seen classes. Additionally, we introduce a strict-to-relaxed penalty term for enforcing this constraint, which depends on the semantic distance of class c to the other seen classes. Formally, the Semantic Neighbourhood loss is given by

$$L_{S_n} = \sum_{x_i^{c;d} \in \mathcal{D}_{\text{train}}} \sum_{j \in \mathcal{C}_{\text{train}}} w(c) \|D(f_i^{c;d}) - D_{\text{gt}}(f_i^{c;d})\|^2 \quad (1)$$

where $D(f_i^{c;d}) \in \mathbb{R}^{|\mathcal{C}_{\text{train}}|}$, such that, its j^{th} element contains the Euclidean distance between $f_i^{c;d}$ and the semantic-information of the j^{th} class, denoted by a^j . Similarly, the j^{th} element of the corresponding ground-truth distance-vector ($D_{\text{gt}}(f_i^{c;d})$) is the Euclidean distance between a^c and a^j . \odot represents element-wise multiplication. The entry of the weight vector $w(c) \in \mathbb{R}^{|\mathcal{C}_{\text{train}}|}$ is

$$w(c)_j = \exp(-D_n(a^c; a^j)) \quad (2)$$

where, $D_n(a^c; a^j) = \frac{D(a^c; a^j)}{\max_{k:k \in \mathcal{C}_{\text{train}}} D(a^c; a^k)}$; $D(;;)$ represents the Euclidean distance between two vectors a and b is an experimental hyper-parameter, set using validation set accuracy. For a smaller $D_n(a^c; a^j)$ (semantically similar classes), $w(c)_j$ is higher, which enforces greater emphasis to preserve the relative distance for the similar classes as desired. For distant classes, this constraint is less strictly enforced.

4.2. Domain-independent representation

Here, we propose to learn the data-representation by incorporating mix-up based supervision, such that SnMpNet only learns the class-information of the mixed sample and obtains a domain-invariant embedding.

Mixing up input data: Inspired by [17], we mix samples from multiple classes, as well as from multiple domains to form the set $\mathcal{D}_{\text{mixup}} = \{x, g\}$, such that

$$x = x_i^{c;d} + (1 - \beta) [x_j^{p;d} + (1 - \gamma) x_k^{r;n}] \quad (3)$$

where $\beta \sim \text{Beta}(\cdot; \cdot)$ and $\gamma \sim \text{Bernoulli}(\cdot)$, and β, γ being hyper-parameters. Clearly, $\beta = 1$ results in intra-domain and $\beta = 0$ results in cross-domain mix-up. We use the samples $\mathcal{D}_{\text{mixup}}$ for training.

Mixture Prediction Loss: We aim to remove any domain-related information from $g = \text{bb}(x)$ by means of such cross-domain mixture samples. For this, we propose a novel Mixture Prediction loss where the network is trained

Figure 1: Depiction of the proposed SHMPNet (a) illustrates the training methodology using mix-up and customized Mixture Prediction layer and Semantic Neighbourhood layer on top of the Base Model (b) illustrates the testing under proposed UCDR protocol, where query samples during retrieval can come from seen domain and unseen category

to predict the exact proportion of the component categories indices and zero at other places, e.g., for generated through in a sample \mathbf{x} , and forgets about its component domains. $\mathbf{t} = 1$ in (3):

We incorporate this constraint before learning the space representation, to ensure that the Semantic Neighbourhood loss does not get hindered by any domain knowledge.

For this, we compute the logit-vectors \mathbf{g} by passing it through the mixture prediction layer. The softmax activation on the j^{th} element of $\mathbf{M}_p(\mathbf{g})$ can be interpreted as the probability that sample \mathbf{x} belongs to class j :

$$\text{Prob}(\mathbf{x} \in \text{Class } j) = \frac{\exp(\mathbf{M}_p(\mathbf{g})_j)}{\sum_{t \in \mathcal{C}_{\text{train}}} \exp(\mathbf{M}_p(\mathbf{g})_t)} \quad (4)$$

where $\mathbf{M}_p(\mathbf{g})_j$ is the logit-score obtained at j^{th} index. However, \mathbf{g} contains characteristics from its component classes, according to the mixing coefficients. We propose to predict the mixing coefficients of those component classes through a soft cross-entropy loss designed as:

$$L_{\text{Mp}} = \sum_{\mathbf{x} \in \mathcal{D}_{\text{mixup}}} \sum_{t=1}^{j \in \mathcal{C}_{\text{train}}} \mathbf{t}_t \log \text{Prob}(\mathbf{x} \in \text{Class } t) \quad (5)$$

where, \mathbf{t}_t is t^{th} -element of $j \in \mathcal{C}_{\text{train}}$ j -dimensional vector, with the mixing coefficients at their corresponding class-

$$\mathbf{t}_t = \begin{cases} 1 & \text{if } t = c \\ 0 & \text{if } t = p \\ 0 & \text{otherwise} \end{cases}$$

Thus, the network only remembers the cross-category mixing proportions and is independent of the input domains.

Mix-up Classification loss: Finally, we impose the standard cross-entropy loss to ensure that the discriminability of feature-space is preserved. For this, we classify the latent \mathbf{z} -space representation $\mathbf{f} \in \mathbb{R}^d$ of the mixed-up sample \mathbf{x} into its component classes as in [33][17]. Additionally, we also want to maintain a meaningful semantic structure of the latent space, to make provision for unseen classes. Here, we utilize the semantic information of the seen classes to address both the stated requirements. Towards this goal, we compute the logit-score for feature $\mathbf{ass}(\mathbf{f}) \in \mathbb{R}^{|\mathcal{C}_{\text{train}}|}$, so that its j^{th} element can be expressed as

$$s(\mathbf{f})_j = \frac{\exp(\text{cosine-similarity}(\mathbf{f}, \mathbf{a}^j))}{\sum_{t \in \mathcal{C}_{\text{train}}} \exp(\text{cosine-similarity}(\mathbf{f}, \mathbf{a}^t))} \quad (6)$$

Similar to (4), $s(f)_j$ also represents the probability that belongs to j^{th} -class. Now, if x came from a particular class, classification can be done by minimizing the following

$$L_{CE}(y(x); s(f)) = \sum_{x \in 2D_{mixup}} \sum_{j \in C_{train}} y(x)_j \log(s(f)_j) \quad (7)$$

where, $y(x)_t$ is the t^{th} element of the one-hot representation of x 's ground-truth class. Since the input does not belong to a single class, we cannot directly use such computation. Instead, we extend equation (7) to accommodate all the component classes x as follows,

$$L_{CE}^{mix} = L_{CE}(y(x_i^{c:d}); s(f)) + (1 - \alpha) L_{CE}([y(x_j^{p:d}) + (1 - \alpha) y(x_k^{r:n})]; s(f)) \quad (8)$$

4.3. Combined Loss Function

Finally, to account for both unseen classes and unseen domains during retrieval, we combine the advantages of both the above representations seamlessly in the proposed framework. To incorporate the effect of inter-class and inter-domain mix-up in the Semantic Neighbourhood loss, we compute $D(f)$ and $D_{gt}(f)$ instead of $D(f_i^{c:d})$ and $D_{gt}(f_i^{c:d})$ in (1). In addition, we evaluate the mixed-up semantic information of x as the combination of its component classes in appropriate ratio as,

$$a = a^c + (1 - \alpha)[a^p + (1 - \alpha)a^r] \quad (9)$$

This modification reflects in evaluation of $D_{gt}(f)$ as its j^{th} component becomes the Euclidean distance between a^j . With this modification, the mix-up based supervision is introduced not only at the image-level, but also in the semantic information level. Combining all the loss components, the final loss to train the model is

$$L = L_{CE}^{mix} + \lambda_1 L_{Mp} + \lambda_2 L_{Sn} \quad (10)$$

where λ_1 and λ_2 are experimental hyper-parameters to balance the contribution of the different loss components.

4.4. Retrieval

During retrieval, for any query data q , we extract the latent space representation $q \in \mathbb{R}^m$ using the trained model. Similarly, we also extract the latent representations of the search set samples $s \in \mathbb{R}^m$ as f_s . We use Euclidean distance between q and f_s ; $s = 1; \dots; |D_s|$ to rank the search set images in the final retrieval list.

Implementation Details: We use PyTorch 1.1.0 and

a single GeForce RTX 2080 Ti GPU for implementation. Models are trained for a maximum of 100 epochs with early stopping of 15 epochs based on the validation set performance. We use SGD with nesterov momentum 0.9 and a batch size of 60 to solve the optimization problem, with an initial learning rate of 1e-3, decayed exponentially to 1e-6 in 20 epochs. 300-d GloVe [22]-embeddings and L2-normalized word2vec embeddings (300-d) [17] are used as the semantic information for Sketchy and DomainNet respectively. The key hyperparameters of SnMpNet are α , λ_1 and λ_2 , which are set as 2e-1; 2e-1, and $\lambda_2 = 1$ for both the datasets.

5. Experiments

We now present the experimental evaluation for the proposed SnMpNet. To the best of our knowledge, this is the first work addressing UCDR, thus there are no established baselines for direct comparison. First, we analyze SnMpNet for UCDR protocol, where only the classes are unseen during retrieval. We specifically consider the application of ZS-SBIR, which is well-explored in literature, to directly compare SnMpNet with current SOTA in ZS-SBIR. Then, we extend our evaluation to the completely general UCDR setting. We start with a brief introduction to the datasets.

Datasets Used: We use two datasets for the experiments. Sketchy extended [24] contains 75,471 sketches and 73,002 images from 125 categories and is used for UCDR. To obtain completely unseen test classes (if pre-trained backbones are used), we follow the split in [32] and consider 21-classes (not part of ImageNet-1K) to be unseen. Among rest 104 seen classes, following [4], 93 and 11-classes are used for training and validation respectively. DomainNet [21] has approximately 6,00,000 samples from 345 categories, collected in six domains, namely, Clip-art, Sketch, Real, Quickdraw, Infograph, and Painting and is used for UCDR and UCDR experiments. Following [17], the test set is formed with 5 unseen classes. Rest 245 and 55 classes are used for training and validation [17]. In addition, we leave one domain (randomly selected) out while training, to create an unseen domain query. The search-set is constructed with real images from seen and/or unseen classes.

5.1. UCDR Evaluation

We first analyze SnMpNet for UCDR, specifically, ZS-SBIR, where the query domain is sketch and the search set contains images, both from a set of classes unseen to the model. Here, we train SnMpNet with sketch and image data, following the same training protocol as in [4][32]. Baseline Methods: First, we discuss the baseline methods and their modifications used for fair comparison. Specifically, we develop variants with no access to domain-label

| Method | | Backbone network | output dim. | mAP@200 | Prec@200 |
|---------------------------|-----------------------------------|------------------|-------------|---------|----------|
| Existing SOTA | CVAE [32] (ECCV, 2018) | VGG-16 | 1024 | 0.225 | 0.333 |
| | Doodle-to-Search [4] (CVPR, 2019) | VGG-16 | 300 | 0.4606 | 0.3704 |
| | SAKE-512 [15] (ICCV, 2019) | SE-ResNet50 | 512 | 0.497 | 0.598 |
| | SAKE-512 (our evaluation) | SE-ResNet50 | 512 | 0.6246 | 0.5518 |
| Doodle-to-search variants | Doodle-SingleNet | VGG-16 | 300 | 0.3743 | 0.3308 |
| | Doodle-SingleNet-w/o Label* | VGG-16 | 300 | 0.3726 | 0.3233 |
| | Doodle-SE-SingleNet | SE-ResNet50 | 300 | 0.4022 | 0.3595 |
| | Doodle-SE-SingleNet-w/o Label* | SE-ResNet50 | 300 | 0.3980 | 0.3508 |
| SAKE-variants | SAKE-512-w/o Label* | SE-ResNet50 | 512 | 0.5484 | 0.4880 |
| | SAKE-300-w/o Label* | SE-ResNet50 | 300 | 0.5192 | 0.4605 |
| SnMpNet | | SE-ResNet50 | 300 | 0.5781 | 0.5155 |

Table 1: Comparison for ZS-SBIR on Sketchy extended [32]. Methods marked with '*' can potentially be used for UCDR.

| Training Domains | Query Domain | Method | Unseen class Search Set | | Seen+Unseen class Search Set | |
|--|--------------|------------------|------------------------------------|----------|---|----------|
| | | | mAP@200 | Prec@200 | mAP@200 | Prec@200 |
| Real Quickdraw Infograph Painting Clip-art | Sketch | EISNet-retrieval | 0.2611 | 0.2061 | 0.2286 | 0.1805 |
| | | CuMix-retrieval | 0.2736 | 0.2168 | 0.2428 | 0.1935 |
| | | SnMpNet | 0.3007 | 0.2432 | 0.2624 | 0.2134 |
| Real Sketch Infograph Painting Clip-art | Quickdraw | EISNet-retrieval | 0.1273 | 0.1016 | 0.1101 | 0.0870 |
| | | CuMix-retrieval | 0.1304 | 0.1006 | 0.1118 | 0.0852 |
| | | SnMpNet | 0.1736 | 0.1284 | 0.1512 | 0.1111 |
| Real Sketch Infograph Quickdraw Clip-art | Painting | EISNet-retrieval | 0.3599 | 0.2913 | 0.3280 | 0.2653 |
| | | CuMix-retrieval | 0.3710 | 0.3001 | 0.3400 | 0.2751 |
| | | SnMpNet | 0.4031 | 0.3332 | 0.3635 | 0.3019 |
| Real Sketch Painting Quickdraw Clip-art | Infograph | EISNet-retrieval | 0.1878 | 0.1512 | 0.1658 | 0.1323 |
| | | CuMix-retrieval | 0.1931 | 0.1543 | 0.1711 | 0.1361 |
| | | SnMpNet | 0.2079 | 0.1717 | 0.1800 | 0.1496 |
| Real Sketch Painting Quickdraw Infograph | Clip-art | EISNet-retrieval | 0.3585 | 0.2792 | 0.3251 | 0.2496 |
| | | CuMix-retrieval | 0.3764 | 0.2911 | 0.3428 | 0.2627 |
| | | SnMpNet | 0.4198 | 0.3323 | 0.3765 | 0.2959 |
| Average | | EISNet-retrieval | 0.2589 | 0.2059 | 0.2315 | 0.1829 |
| | | CuMix-retrieval | 0.2689 | 0.2126 | 0.2417 | 0.1905 |
| | | SnMpNet | 0.3010 | 0.2418 | 0.2667 | 0.2144 |

Table 2: UCDR evaluation results on DomainNet for two different scenarios, when the search set contains (1) ~~only~~ class image samples, and (2) ~~both~~ ~~seen~~ and unseen class samples.

| Training Domains | Query Domain | Method | Unseen class Search Set | | Seen+Unseen class Search Set | |
|--|--------------|------------------|------------------------------------|----------|---|----------|
| | | | mAP@200 | Prec@200 | mAP@200 | Prec@200 |
| Real Quickdraw Infograph Painting Clip-art | QuickDraw | EISNet-retrieval | 0.2475 | 0.1906 | 0.2118 | 0.1627 |
| | | CuMix-retrieval | 0.2546 | 0.1967 | 0.2177 | 0.1699 |
| | | SnMpNet | 0.2888 | 0.2314 | 0.2366 | 0.1918 |
| Real Sketch Infograph Painting Clip-art | Sketch | EISNet-retrieval | 0.3719 | 0.3136 | 0.3355 | 0.2822 |
| | | CuMix-retrieval | 0.3689 | 0.3069 | 0.3300 | 0.2714 |
| | | SnMpNet | 0.4221 | 0.3496 | 0.3767 | 0.3109 |

Table 3: UCDR-evaluation results on DomainNet for two different scenarios, when the search set contains (1) ~~only~~ class image samples, and (2) ~~both~~ ~~seen~~ and unseen class samples.

of data, so that they can handle unseen domain query data.

1. Doodle-to-Search [4] trains two parallel VGG-16 networks with a triplet loss to generate the final embedding for retrieval. We develop the following variants of this network as:
 - Doodle-SingleNet. We replace the architecture in [4] with a single branch of VGG-16, which can take data from any domain as input.
 - Doodle-SingleNet-w/o Label. We further modify Doodle-SingleNet and remove the domain-discriminator loss function from the training process,

so that it can be applied to any unseen domain data.

- For fair comparison, we replace the VGG-16 backbone in both these variants with SE-ResNet50 and develop Doodle-SE-SingleNet and Doodle-SE-SingleNet-w/o Label respectively.
2. SAKE [15] has a single branch of network, with SE-ResNet50 as backbone. It processes both sketch and image data, augmented with a binary domain-label, and knowledge transfer from a pre-trained teacher network. For comparing with SnMpNet, we develop the following variant of SAKE:

- SAKE-w/o Label. In this variant, we remove the binary domain-indicator from the training process.
- As in SAKE [15], we perform experiments of this variant with embeddings of different dimensions.

Apart from Doodle-to-search [4] and SAKE [15], we have also compared SnMpNet with CVAE [32]. We summarize the comparisons in terms of mAP@200 and Prec@200 in Table 1. We observe the best performance is obtained through the SAKE-model with the domain indicator (our evaluation). Note that this model cannot be used for unseen query domains in UCDR protocol, because of the domain indicator. We also observe that the performance of both the state-of-the-art approaches, Doodle-to-Search [4] and SAKE [15] degrade drastically when either the domain-specific two-branch architecture or the domain-indicator is removed. SnMpNet outperforms these variants, CVAE and Doodle-to-Search, which justifies its effectiveness.

5.2. UCDR Evaluation

We now extend our evaluation for the fully generalized UCDR protocol on DomainNet [21]. Since there is no existing baseline for this, we develop two variants of very closely related works present in literature. We start with a brief description of these variants.

Baseline Methods: We consider two state-of-the-art approaches developed for related applications, namely 1) EISNet [28], which is the SOTA for DG and 2) CuMix [17], the first work to address ZSDG. Since these have been developed for classification, we make minimal changes in the networks, to address the retrieval task in UCDR.

1) EISNet-Retrieval: We incorporate a 300-d linear-layer in the classification branch in [28], whose output is used as the domain-invariant feature for UCDR.

2) CuMix-Retrieval: For fair comparison, we use SE-ResNet50 as backbone with a 300-d linear layer on top of in [17] and incorporate the image and feature-level mixing method, as proposed in CuMix. We discuss the details of these modifications in Supplementary.

For UCDR, we train with seen class samples from 5-domains, leaving one domain out. The seen class samples from this unseen domain are used as query for evaluation. We evaluate for two configurations of search set, where it contains images from: (a) only unseen classes, and (b) both seen and unseen classes. Clearly, (b) is more challenging than (a), because of the scope of greater confusion. We report the individual results on all 5-domains (except Data) as query, as well as the average retrieval accuracy in Table 2. We observe that for all the approaches, the performance degrades significantly when both seen and unseen classes are present in the search set. However, SnMpNet outperforms the other baselines by a considerable margin.

¹SAKE-model trained using the split and evaluation in [4]

6. Analysis

We analyze the contribution of different components of SnMpNet, and its performance in other retrieval scenarios.

UCDR-Evaluation on DomainNet: We now present the evaluation of SnMpNet for UCDR on DomainNet. Here the query domain Sketchor QuickDraw is seen, but the query samples belong to unseen classes. We perform experiments for two configurations of search set as UCDR. From Table 3, we observe that SnMpNet outperforms the other approaches.

UCDR-Evaluation on DomainNet: Here, for completeness, we evaluate SnMpNet for UCDR, where the query belongs to a seen class, but from unseen domain. To construct the query set, we randomly select 25% samples from each of the seen classes from Sketch domain. The search set contains images from seen classes. From Table 4, we observe that SnMpNet significantly outperforms the two strong baselines. This setup can be considered as a modification of domain generalization problem for the retrieval task. Here, the knowledge gap between the training and test classes is not present. The main challenge for the retrieval model is to address the domain gap because of the unseen query domain.

| Method | mAP@200 | Prec@200 |
|------------------|---------|----------|
| EISNet-retrieval | 0.2210 | 0.1094 |
| CuMix-retrieval | 0.2703 | 0.1224 |
| SnMpNet | 0.3529 | 0.1657 |

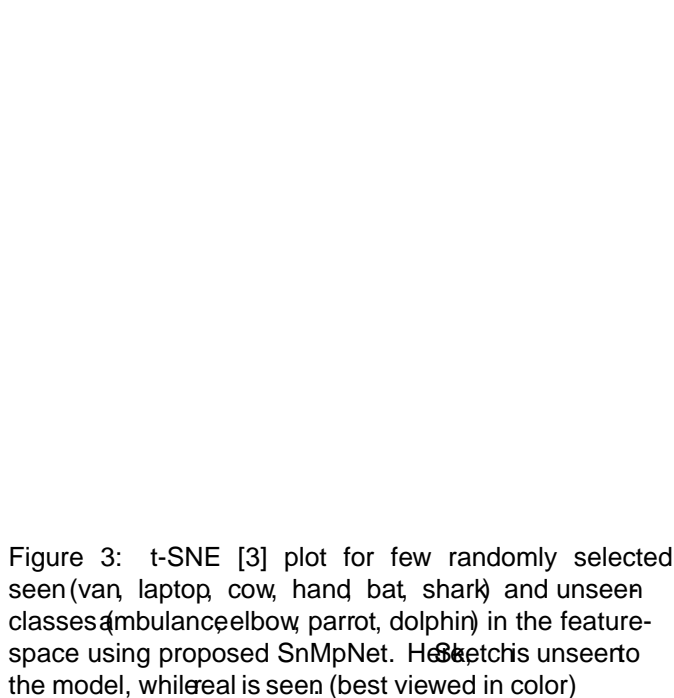
Table 4: UCDR-evaluation on DomainNet for unseen sketch query domain. The search set contains only seen class real images. The models are trained on 5-domains Real QuickDraw, Infograph, Painting and Clip-art.

Ablation Study: Now, we analyze the effectiveness of different components of SnMpNet on Sketchy-extended for ZS-SBIR. We first consider the simplest form of our network (Base N/W), which is SE-ResNet50, trained with the cross-entropy loss evaluated as the cosine-similarity of model output with seen-classes' semantic information [22] (equation (7)). The performance of Base N/W, and as each loss component is appended to the base module is summarized in Table 5. We observe that each component contributes positively to the overall performance.

| Proposed Network Variants | mAP@200 | Prec@200 |
|------------------------------------|---------|----------|
| Base N/W | 0.5218 | 0.4497 |
| Base N/W + L_{S_n} ($= 0$) | 0.5593 | 0.5002 |
| Base N/W + L_{S_n} ($= 2$) | 0.5613 | 0.5030 |
| Base N/W + L_{CE}^{mix} | 0.5252 | 0.4530 |
| Base N/W + $L_{CE}^{mix} + L_{Mp}$ | 0.5665 | 0.4989 |
| SnMpNet | 0.5781 | 0.5155 |

Table 5: Ablation study of the proposed SnMpNet framework for ZS-SBIR on Sketchy extended dataset [32].

Figure 2: Top-8 retrieved Images for UCDR and QCDR protocols on DomainNet with QuickDraw being the unseen query domain. Same query is considered for both the search set configurations. Green and Red borders indicate correct and incorrect retrievals respectively. (best viewed in color)



6.1. Qualitative Results

Figure 2 shows top-8 retrieved images for few queries for UCDR and QCDR, with QuickDraw as unseen domain. As expected, the results degrade when both seen and unseen classes are present in the search set. We also observe that some of the incorrect retrievals are because of shape simi-

larities between classes, like helicopter and windmill, while some others are due to co-occurrence of different classes in the same images, sweater and elbow for skateboard.

The t-SNE [3] plot of the feature-space for some randomly selected classes from seen (image) and unseen (sketch) domains is shown in Figure 3. We observe that the unseen classes - namely ambulance, dolphin, parrot, and elbow from the unseen domain sketch are placed in the neighbourhood of the related seen classes - van, shark, bat, and hand respectively from seen domain, image, further justifying the effectiveness of the proposed framework.

7. Conclusion

In this work, we proposed a novel framework SnMpNet for the Universal Cross-domain Retrieval task. To the best of our knowledge, this is the first work which can handle query data from unseen classes and unseen domains for retrieval. In addition to defining the experimental protocol, we also proposed a novel framework, SnMpNet, which introduces two novel losses Semantic Neighbourhood loss and Mixture Prediction loss for the UCDR task. Extensive experiments and comparisons on two large-scale datasets, corroborate effectiveness of the proposed SnMpNet.

Acknowledgements This work is partly supported through a research grant from SERB, Department of Science and Technology, Govt. of India.

References

- [1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmidt. Label embedding for attribute-based classification, 2013. CVPR.
- [2] F. M. Carlucci, A. Delnciente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles, 2019. CVPR.
- [3] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008.
- [4] S. Dey, P. Riba, A. Dutta, and J. Lladós. Doodle-to-search: practical zero-shot sketch-based image retrieval, 2019. CVPR.
- [5] A. Dutta and Z. Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval, 2019. CVPR.
- [6] T. Dutta, A. Singh, and S. Biswas. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir, 2020. ECCV.
- [7] T. Dutta, A. Singh, and S. Biswas. Styleguide: zero-shot sketch-based image retrieval using style-guided image generation. *IEEE T-MM*, 2020.
- [8] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task auto-encoders, 2015. ICCV.
- [9] D. Li, Y. Yang, Y. Z. Song, and T. M. Hospedales. Learning to generalize: meta-learning for domain generalization, 2018. AAAI.
- [10] D. Li, J. Zhang, Y. Yang, C. Liu, Y. Z. Song, and T. M. Hospedales. Episodic training for domain generalization, 2019. ICCV.
- [11] H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization via adversarial feature learning, 2018. CVPR.
- [12] S. Li, D. Chen, B. Liu, N. Yu, and R. Zhao. Memory-based neighbourhood embedding for visual recognition, 2019. ICCV.
- [13] Y. Li, Y. Yang, W. Zhou, and T. Hospedales. Feature-critic networks for heterogenous domain generalization, 2019. ICML.
- [14] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: fast free-hand sketch-based image retrieval, 2017. CVPR.
- [15] Q. Liu, L. Xie, H. Wang, and A. Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval, 2019. ICCV.
- [16] P. Lu, G. Huang, Y. Fu, G. Guo, and H. Lin. Learning large euclidean margin for sketch-based image retrieval, 2018. <https://arxiv.org/abs/1812.04275v1>.
- [17] M. Mancini, Z. Akata, E. Ricci, and B. Caputo. Towards recognizing unseen categories in unseen domains, 2020. ECCV.
- [18] U. Maniyar, J. KJ, A. A. Deshmukh, U. Dogan, and V. Balasubramanian. Zero-shot domain generalization, 2020. BMVC.
- [19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. <https://arxiv.org/abs/1301.3781v3>.
- [20] K. Muandet, D. Balduzzi, and B. Scholkopf. Domain generalization via invariant feature representation, 2013. ICML.
- [21] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation, 2019. ICCV.
- [22] J. Pennington, R. Socher, and C. D. Manning. Glove: global vectors for word representation, 2014. EMNLP.
- [23] B. Romera-Paredes and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning, 2015. ICML.
- [24] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 35(4):1–12, 2016.
- [25] E. Schonfeld, S. Emrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero-shot and few-shot learning via aligned variational autoencoders, 2019. CVPR.
- [26] Y. Shen, L. Liu, F. Shen, and L. Shao. Zero-shot sketch-image hashing, 2018. CVPR.
- [27] W. Thong, P. Mettes, and C. G. M. Snoek. Open cross-domain visual search. *CVIU*, 200, 2020.
- [28] S. Wang, L. Yu, C. Li, C. W. Fu, and P. A. Heng. Learning from extrinsic and intrinsic supervisions for domain generalization, 2020. ECCV.
- [29] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning: a comprehensive evaluation of the good, the bad and the ugly. *IEEE T-PAMI*, 41(9):2251–2265, 2018.
- [30] G. S. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, and L. Shao. Region graph embedding network for zero-shot learning, 2020. ECCV.
- [31] X. Xu, M. Yang, Y. Yang, and H. Wang. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval, 2020. IJCAI.
- [32] S. K. Yelamarthy, S. K. Reddy, A. Mishra, and A. Mittal. A zero-shot framework for sketch-based image retrieval, 2018. ECCV.
- [33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: beyond empirical risk minimization, 2018. ICLR.
- [34] R. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. Tao-Shen, and L. Van Gool. Generative domain-migration hashing for sketch-to-image retrieval, 2018. ECCV.
- [35] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding, 2016. CVPR.