Open Cross-Domain Visual Search

William Thong  Pascal Mettes  Cees G.M. Snoek
University of Amsterdam
{w.e.thong,p.s.m.mettes, cgmsnoek}@uva.nl

Abstract

This paper introduces open cross-domain visual search, where categories in any target domain are retrieved based on queries from any source domain. Current works usually tackle cross-domain visual search as a domain adaptation problem. This limits the search to a closed setting, with one fixed source domain and one fixed target domain. To make the step towards an open setting where multiple visual domains are available, we introduce a simple yet effective approach. We formulate the search as one of mapping examples from every visual domain to a common semantic space, where categories are represented by hyperspherical prototypes. Cross-domain search is then performed by searching in the common space, regardless of which domains are used as source or target. Having separate mappings for every domain allows us to search in an open setting, and to incrementally add new domains over time without retraining existing mapping functions. Experimentally, we show our capability to perform open cross-domain visual search. Our approach is competitive with respect to traditional closed settings, where we obtain state-of-the-art results on six benchmarks for three sketch-based search tasks.

1. Introduction

This work investigates categorical search in an open cross-domain setting. Cross-domain visual search has made a lot of progress recently in a closed setting, where natural images [13, 42] or 3D shapes [22–24] examples of categories are retrieved from sketches. Such a closed setting only considers a search from one fixed source domain to one fixed target domain. In practice however, categories come in many forms [25, 36, 51]. Hence, we may have queries from several source domains, or want to search in any possible combinations of source and target domains. As a first contribution of this work, we introduce open cross-domain visual search: we search for categories from any source domain to any target domain, with the ability to search from and within multiple domains simultaneously.

Figure 1: Open cross-domain visual search. We search for categories from any number of source domains to any number of target domains. Mapping examples to a common semantic space enables any possible combinations of cross-domain search, without hassle.

Addressing the domain gap between source and target domains has proven to be effective for cross-domain search [4, 10, 12, 13, 44, 48, 52, 54]. While intuitive and compelling, focusing on domain adaptation with pair-wise training makes the search unsuited for an open setting where multiple visual domains are available. To move towards an open setting, we should align examples by the very thing that unites them, namely their semantics, rather than aligning the domains they originate from. As a second contribution, we propose a simple approach for open cross-domain visual search, where we start from a common semantic space in which categories are represented by hyperspherical prototypes. For every domain, we learn a function to map visual inputs to their corresponding prototypes in the common semantic space, as illustrated in Figure 1. Query representations for search are further refined with neighbours from other domains through a spherical linear interpolation operation. Once trained, the proposed formulation allows us to search among any pair of domains. Since all domains are now aligned semantically in the common semantic space, this enables a search from multiple source domains or in multiple target domains. Lastly, new domains can be added on-the-fly, without the need to retrain previous models.

As a third contribution, we perform extensive evaluations to demonstrate our ability to perform open cross-domain visual search, as well as our efficacy in stan-
standard closed settings compared to current approaches. For open cross-domain visual search, we perform several novel demonstrations showing: i) a search between any pair of source and target domains without hassle, ii) a search from multiple source domains, and iii) a search in multiple target domains. While designed for the open cross-domain setting, we find that our approach works well in conventional closed settings also. We compare on sketch-based image and 3D shape retrieval. Across three tasks and six benchmarks, we obtain state-of-the-art results, highlighting the effectiveness of our approach. All code and setups will be released to foster further research in open cross-domain visual search.

2. Related Work

Cross-domain visual search A wide range of works have focused on cross-domain visual search by setting the source domain as sketches. Natural images [13, 42] or 3D shapes [19, 22, 24] of the same category are then retrieved given the sketch query. When searching for natural images from a sketch, a common approach is to bridge the domain gap between sketches and images [10, 12, 44, 54]. Shen et al. [44] fuse sketch and image representations with a Kronecker product layer [20], while Yelamarthi et al. [54] introduce domain confusion with generative models. Dey et al. [10] combine gradient reversal layers [15] with metric learning losses [16, 43] to further enforce a domain agnostic embedding space. Dutta and Akata [12] tie the semantic space with visual features by learning to generate them. Alternatively, Liu et al. [27] preserve the knowledge from a pre-trained model. Hu et al. [19] have also explored few-shot image classification by synthesizing classifiers from sketches. By focusing on domain adaptation, current approaches are limited to a mapping from one source domain (e.g., sketch) to one target domain (e.g., image). In this paper, we move this paradigm towards open cross-domain visual search, where search occurs from any source domain to any other target domain.

Searching for 3D shapes from a sketch has been accelerated by the SHREC challenges [22–24]. A recent trend is to perform cross-domain retrieval from 2D image domain to the 3D shape domain [4, 8, 40, 47, 48, 52]. In this setting, Wang et al. [48] map both sketches and 3D shapes in a similar feature space with a Siamese network [6, 17], while Tasse et al. [47] learn to regress to a semantic space with a ranking loss [14]. Dai et al. [8] correlate both sketch and 3D shape representations to bridge the domain gap. Xie et al. [52] employ the Wasserstein distance to create a barycentric representation of 3D shapes. Qi et al. [40] apply loss functions on the label space rather than the feature space. Chen et al. [4] propose an advanced sampling of 2D views for unaligned shapes. Akin to [40, 47], we place a central role on semantics for cross-domain search. In this paper, we go beyond searching for only 3D shapes to searching among any number of available target domains. We map multiple domains to semantic prototypes in a common embedding space, which alleviates the need for multi-stage training and negative sampling schemes.

Using multiple domains has recently been investigated in unsupervised domain adaptation [7, 37] or domain generalization [1] works, where the task is to classify unlabeled target samples by learning a classifier on labeled source samples. Learning a classifier from multiple sources has been shown to be beneficial for both tasks (e.g., [2, 11, 36, 53, 57]). In this paper, we focus on a different multi-domain task, namely open cross-domain visual search.

Learning with prototypes Learning metric spaces with prototypes for image retrieval [9, 28, 34, 49, 50, 55] and classification [5, 31, 32] provides a simpler alternative to common contrastive [6, 17] or triplet [43] loss functions. No complex sampling is required, making the training easier in return [34, 50]. A first line of work learns prototype representations, such as the center loss [50], the proxy loss [34, 55], and derivations that introduce a margin in the distance measure [9, 28, 49]. A second line of work fixes prototype representations. Mensink et al. [31] set class means as prototypes in the embedding space for zero-shot classification. Chintala et al. [5] show that regressing to one-hot prototypes is close to a softmax classifier. Mettes et al. [32] better position prototypes on a low-dimensional hypersphere for classification and regression. Here, we take inspiration from this metric learning literature with prototypes and leverage them to the problem of open cross-domain visual search. We create a common semantic space where classes are represented by prototypes on a hypersphere. Every domain has its respective model to map visual inputs to the common semantic space where the cross-domain search occurs.

3. Method

3.1. Problem formulation

The problem formulation of open cross-domain visual search is illustrated in Figure 2. While the closed cross-domain setting focuses on one fixed source s and one fixed target t, the open cross-domain setting searches for categories from any source domain s_k to any target domain t_k. As multiple domains now become available, this opens the door for combining multiple domains at both source and target positions. Thus, the main difference between the closed setting and the open setting lies in the ability to leverage multiple domains for categorical cross-domain search.

Formally, let D denote the set of all domains to be considered. Rather than making an explicit split of a dataset into source and target, we consider a large combined visual collection \( T = \{(x_n^d, y_n)\}_{n=1}^N \), where \( x_n^d \in \mathcal{I}_d \) denotes an input example from a visual domain \( d \in D \) of category
y_n \in \mathcal{Y}$. In other words, \( \mathcal{Y} \) is common and shared among all domains \( \mathcal{D} \) but is depicted differently from domain \( d_i \) to domain \( d_j \), with \( i \neq j \).

Categorical search consists in using a sample query \( x_i \) from domain \( d_i \) to retrieve samples of the same category \( y \) in the gallery of domain \( d_j \). If \( i \neq j \), this corresponds to a cross-domain categorical search as the search occurs across two different domains. A closed setting only considers \( |\mathcal{D}| = 2 \), i.e. with only a single source domain and a single target domain. We define the open setting as comprising \( |\mathcal{D}| > 2 \). This stimulates novel search configurations. For example, we may want to combine two queries \( x_i, x_j \) of two different domains \( i \neq j \) to search in the gallery of a third domain \( k \). Conversely, given a sample query \( x_i \), we can search in the combined gallery of multiple domains.

### 3.2. Proposed approach

We pose open domain visual search as projecting any number of heterogeneous domains to prototypes on a common and shared hyperspherical semantic space. First, we outline how to represent categories in the semantic embedding space. Second, we propose a mapping function for every domain to the common semantic embedding space.

#### Categorical prototypes.

We leverage the concept of prototypes to represent categories in a common semantic space. Every category is represented by a unique real-valued vector, corresponding to a categorical prototype. Hence, the objective is to align examples, coming from different domains but with the same category label, to the same categorical prototype in the common semantic space. For every category \( y \in \mathcal{Y} \), we denote its prototype on the semantic space as \( \phi(y) \in \mathbb{S}^{D-1} \) for a \( D \)-dimensional hypersphere. Relying on semantic relations enables to search for unseen classes using models trained on seen categories \([14,35]\). In this work, we opt for word embeddings (e.g. word2vec \([33]\) or GloVe \([38]\)) to represent categories, as these embeddings adhere to the semantic relation property.

#### Mapping domains to categories.

For every domain \( d \in \mathcal{D} \), we learn a separate mapping function \( f_d(\cdot) \in \mathbb{S}^{D-1} \) to the common and shared semantic space. Separate mapping functions are not only easy to train, they also enable us to incorporate new domains over time. Indeed, we only have to train the mapping of the new incoming domain without retraining previous mapping functions of existing domains. The mapping function itself is formulated as a convolutional network (ConvNet) with \( \ell_2 \)-normalization on the \( D \)-dimensional network outputs.

We propose the following function to map an example of a certain domain \( d \) to its categorical prototype in the shared semantic space:

\[
p(y|x^d, d) = \frac{\exp \left( -s \cdot c(f_d(x^d), \phi(y)) \right)}{\sum_{y' \in \mathcal{Y}} \exp \left( -s \cdot c(f_d(x^d), \phi(y')) \right)},
\]

where \( s \in \mathbb{R}_{>0} \) denotes a scaling factor, inversely equivalent to the temperature \([18]\). Intuitively, the scaling controls how samples are spread around categorical prototypes. The amount of scaling \( s \) is a hyperparameter that we study in the supplementary materials. The distance function \( c(\cdot, \cdot) \) is defined as the cosine distance:

\[
c(f_d(x^d), \phi(y)) = 1 - \langle f_d(x^d), \phi(y) \rangle,
\]

where \( \langle \cdot, \cdot \rangle \) is the dot product. As both \( f_d(x) \) and \( \phi(y) \) lie on a hypersphere, they have a unit norm. Finally, learning every mapping function \( f_d \) is done by minimizing the cross-entropy over the training set:

\[
\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \log p(y_n|x_n^d, d).
\]
Given one or more queries from different source domains, we first project all queries to the shared semantic space and average their positions into a single vector. Then, we compute the distance to all target examples to rank them with respect to the query. We can on-the-fly combine source domains to search from or target domains to search within.

3.3. Refining queries across domains

With our approach, a source query is close to target examples from the same category, regardless of the domains of the query and target examples. In practice, inherent variability in the hyperspherical semantic space can cause noise in the similarity measures. We therefore propose to refine the initial query representation using a nearby example from the target domain. Figure 3 illustrates the refinement.

We refine the query representation \( p_0 \) by performing a spherical linear interpolation with a relevant representation \( p_1 \). This relevant representation is either the nearest neighbour in the target set (for retrieval) or the word embedding of the category (for classification). The refined representation operation \( \tilde{p} \) is expressed as:

\[
\tilde{p}(p_0, p_1 | \lambda) = \frac{\sin ((1 - \lambda)\Omega)}{\sin \Omega} p_0 + \frac{\sin (\lambda\Omega)}{\sin \Omega} p_1,
\]

where \( \Omega = \arccos(p_0 \cdot p_1) \) and \( \lambda \in [0, 1] \) controls the amount of mixture in the refinement process. The higher the value of lambda is, the further away the refined representation is from the original representation \( p_0 \). Intuitively, the refinement performs a weighted signal averaging to reduce the noise present in the initial representation. The amount of interpolation \( \lambda \) is a hyperparameter that we study in the supplementary materials.

4. Open Cross-Domain Visual Search

In the first set of experiments, we demonstrate our newly gained ability to perform open cross-domain visual search in three ways. We note that this is a new setting, making direct comparisons to existing works infeasible. First, we demonstrate how we can search from any source to any target domain (mAP@all). Our approach can perform all 72 cross-domain search tasks with ease, as we bypass the need to align domains.

Figure 4: Demonstration 1 for visual search from any source to any target domain (mAP@all). We evaluate on the recently introduced DomainNet [36], which contains 596,006 images from 345 classes. Images are gathered from six visual domains: clipart, infographic, painting, pencil, photo and sketch. We consider retrieval in a zero- and many-shot settings: i) in the zero-shot setting, \( \mathcal{Y} \) is split into \( \mathcal{Y}_{\text{train}} \) and \( \mathcal{Y}_{\text{test}} \), with \( \mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{test}} = \emptyset \), i.e. categories to be searched during inference have not been seen during training; ii) the many-shot setting uses the same categories during both training and testing. The zero-shot setting randomly splits samples into 300 training and 45 testing classes with at least 40 samples per class. The many-shot setting follows the original splits [36]. We report the mean average precision (mAP@all). Briefly, we use SE-ResNet50 [21] pre-trained on ImageNet [41] as a backbone, and word2vec trained on a Google News corpus [33] as the common semantic space. We optimize the loss function with Nesterov momentum [46]. We set the learning rate to 1e−4 with cosine annealing without warm restarts [29] and the batch size to 128.

4.1. From any source to any target domain

First, we demonstrate how searching from any source to any target domain in an open setting is trivially enabled by our approach. Figure 4 shows the result of 72 cross-domain search evaluations; corresponding to all six cross-domain pairs in both the zero-shot and many-shot settings. In our formulation, such an exhaustive evaluation is enabled by training only six models, one for every domain. For comparison, a domain adaptation approach—the standard in current cross-domain search approaches—requires a pair-wise
training of all available domain combinations. Moreover, our formulation allows for an easy integration of new domains, as only the mapping from a new visual domain to the shared semantic space needs to be trained. While approaches based on pair-wise training scale with a quadratic complexity to the number of domains, we scale linearly.

For DomainNet, we find that the photograph domain provides the most effective search whether used as source or target. One reason is the number of available images, which is up to four times larger than other domains. On the other hand, infographs and sketches are very diverse in terms of scale and visual representations, which induces a much more difficult search. We conclude from the first demonstration that search from any source to any target domain is not only feasible with our approach, it can be done easily since we bypass the need to align different domains.

### 4.2. From multiple sources to any target domain

Second, we demonstrate the potential to search from multiple source domains. Due to the generic nature of our approach, we are not restricted to search from a single source. Here, we show that a multi-source search benefits the search in any target domain. For this experiment, we start from the sketch domain as a source and investigate the effect of including queries from the most effective source (photographs) and the least effective source (infographs).

Table 1 highlights the positive effect of searching with an additional domain, rather than a single source domain. When using multiple sources, we simply average the positions in the common semantic space. For fairness, we also evaluate search using two sketches. Across all settings, we find that searching from multiple queries improves relative to using one single sketch query. In the zero-shot setting, including infographs and photographs improves upon sketch-based search only. In the many-shot setting, including infographs improves upon search by one sketch, but not by two sketches, which is not surprising given the low search scores for infographs individually. Including photographs with sketches obtains the highest scores, regardless of the target domain or the evaluation setting.

This demonstration shows the potential of searching from multiple sources. It is better to diversify the search by using multiple domains than include more queries from the same domain. Similar to the first demonstration, this evaluation is a trivial extension to our approach, as we only have to average positions in the shared semantic space, regardless of the domain the examples come from.

### 4.3. From any source to multiple target domains

Third, we demonstrate our ability to search in multiple domains simultaneously. This setting has potential applications for example in untargeted portfolio browsing, where a user may want to explore all possible visual expressions of a category. Exploring in multiple domains also highlights whether certain categories have a preference towards specific domains, which offers an insight on how to best depict those categories. Note that this setting can also be easily extended to include also multiple domains as a source. For the sake of clarity, we use sketch as the source domain and search in the other five domains.
|                   | Sketchy Extended | TU-Berlin Extended |
|-------------------|------------------|--------------------|
|                   | mAP@all prec@100 | mAP@all prec@100   |
| EMS [30]          | n/a              | 0.259 0.369        |
| CAAE [54]         | 0.196 0.284      | n/a                |
| ADS [10]          | 0.369 n/a        | 0.110 n/a          |
| SEM-PCYC [12]     | 0.349 0.463      | 0.297 0.426        |
| SAKE [27]         | 0.535 0.677      | 0.471 0.600        |
| This paper        | **0.649 0.708**  | **0.517 0.557**    |

Table 2: **Comparison 1** to zero-shot sketch-based image retrieval on Sketchy Extended and TU-Berlin Extended (multi-class accuracy). Aligning the semantics, rather than domains, improves cross-domain image retrieval.

5. Closed Cross-Domain Visual Search

Our approach is geared towards open cross-domain visual search, as demonstrated above. To get insight in the effectiveness of our approach for cross-domain visual search in general, we also perform an extensive comparative evaluation on standard cross-domain settings, which search from one source domain to one target domain. In total, we compare on three of the most popular cross-domain search tasks, namely zero-shot sketch-based image retrieval [13, 42, 44], few-shot sketch-based image classification [19], and many-shot sketch-based 3D shape retrieval [22, 24]. For our approach, we simply train one mapping function for the source domain, and one for the target domain using the examples provided during training. Since each approach in closed cross-domain visual search employs different networks and optimizations, an apples-to-apples comparison is not feasible. Hence, we compare our results to the current state-of-the-art results as reported in the respective papers. Implementation details are in the supplementary materials. Below, we handle each comparison separately.

Figure 5 provides qualitative results for six sketches from different categories. We first observe that the results come from multiple target domains, without being explicitly told to do so. We do not need to align results from different target domains, since we measure distance in the common semantic space. For categories such as sun, we have a bias towards retrieving abstract depictions, such as pencil drawings and cliparts, as the sun is a category with a clear abstract representation. Castle on the other hand has a bias towards both distinct cliparts, as well as photographs and paintings. In both cases, all top results are relevant. For categories with more ambiguous sketches, such as river or calculator, retrieved examples resemble the shape of the provided sketch, but do not match the category. Overall, we conclude that searching in multiple domains is not only trivial in our approach, but is also an indicator of the presence of preferential domains for depicting visually categories.

5.1. Zero-shot sketch-based image retrieval

Setup. Zero-shot sketch-based image retrieval focuses on retrieving natural images (target domain) from a sketch query (source domain). We evaluate on two datasets. Sketchy Extended [26, 42] contains 75,481 sketches and 73,002 images from 125 classes. Following Shen et al. [44], we select 100 classes for training and 25 classes for testing. TU-Berlin Extended [13, 56] contains 20,000 sketches and 204,070 images from 250 classes. Similarly, following Shen et al. [44], we select 220 classes for training and 30 classes for testing. For both datasets, we select the same unseen classes as in Liu et al. [27]. Following recent works [12, 27, 44], we report the mAP@all and the precision at 100 (prec@100) scores.

Results. Table 2 compares to five state-of-the-art baselines on both datasets. Baselines mostly focus on bridging the domain gap between sketches and natural images with domain adaptation losses [15, 16]. On Sketchy Extended, our approach outperforms other baselines. On TU-Berlin Extended, we obtain the highest mAP@all, while the recently introduced SAKE by Liu et al. [27] obtains higher prec@100. SAKE is then better at grouping images from the same category together, while our approach is better at retrieving relevant images in the first ranks. We also report on quantized representations in the supplementary materials, with similar improvements over existing baselines. Overall, our formulation based on semantic alignment is competitive with respect to alternatives that focus on domain adaption or knowledge preservation.
Figure 7: Qualitative analysis of few-shot sketch-based image classification. We show the most and least effective sketches for cross-domain classification on Sketchy Extended. Since our approach condenses examples of category to a single prototype in the shared space, we obtain high scores when source sketches are detailed and in canonical views (e.g., cat, mouse, or duck). The accuracy decreases when sketches are drawn badly (e.g., knife), or in non-canonical views (e.g., cat, deer, or car).

Table 3: Comparison 2 to few-shot sketch-based image classification on Sketchy Extended (multi-class accuracy). Our metric learning approach outperforms model regression approaches in few-shot cross-domain classification.

|             | w2v | sketch | image       |
|-------------|-----|--------|-------------|
|             |     | one-shot | five-shot  | one-shot | five-shot |
| M2M [19]    | n/a | 79.93   | n/a        | 93.55    |
| F2M [19]    | 35.90 | 68.16   | 83.01      | 84.12    | 93.89    |
| This paper  | 76.73 | 82.01   | 84.99      | 91.34    | 94.74    |

Table 3: Comparison 2 to few-shot sketch-based image classification on Sketchy Extended (multi-class accuracy). Our metric learning approach outperforms model regression approaches in few-shot cross-domain classification.

Qualitative analysis. To understand how to best employ our approach for few-shot sketch-based image classification, we provide the most and least effective sketches for image classification in Figure 7. Since categories are condensed to a single prototypical sketch, our approach desires sketches with details and in canonical configurations. Results are degraded when such assertions are not met.

5.2. Few-shot sketch-based image classification

Setup. Few-shot sketch-based image classification focuses on classifying natural images from one or a few labeled sketches. The few-shot categories to be evaluated have not been observed during training. Different from the zero-shot retrieval scenario, the few-shot classification setting has access to the labels of the unseen classes in the evaluation phase. For example, this comes through the form of sketches or word embeddings. We report results on the Sketchy Extended dataset [26, 42]. Following Hu et al. [19], we select the same 115 classes for training and 10 classes for testing. We evaluate the performance with the multi-class accuracy and report results over 500 runs. Classification is done by measuring the distance to the class prototypes. We evaluate on three different modes [19]. First, we set the word vectors (w2v) to be prototypes of the unseen classes. Second, we set one or five sketch representations to be prototypes. Third, we use one or five images. The latter is considered as an upper-bound of the cross-domain task.

Results. Table 3 compares our formulation to two baselines introduced by Hu et al. [19]. M2M regresses weights for natural image classification from the weights of the sketch classifier while F2M regresses weights from sketch representations. For the first evaluation mode, we obtain an accuracy of 76.73%, compared to 35.90%, which reiterates the importance of a semantic alignment for categorical cross-domain search. In the few-shot evaluation, we find that the biggest relative improvement is achieved in the one-shot setting. Our approach is then effective for cross-domain classification, especially with a low number of shots.

5.3. Many-shot sketch-based 3D shape retrieval

Setup. Sketch-based 3D shape retrieval focuses on retrieving 3D shape models from a sketch query, where both training and testing samples share the same set of classes. We evaluate on three datasets. SHREC13 [22] is constructed from the TU-Berlin [13] and Princeton Shape Benchmark [45] datasets, resulting in 7,200 sketches and 1,258 3D shapes from 90 classes. The training set contains 50 sketches per class, the testing set 10. SHREC14 [24] contains more 3D shapes and more classes, resulting in 13,680 sketches and 8,987 3D shapes from 171 classes. The training and testing splits of sketches follow the same protocol as SHREC13. We also report on the recently outlined Part-SHREC14 [40], which contains 3,840 sketches and 7,238 3D shapes from 48 classes. The sketch splits also follow the same protocol, while the 3D shapes are now split into 5,812 for training and 1,426 for testing to avoid overlap.

We generate 2D projections for all 3D shape models using the Phong reflection model [39] and render 12 different views by placing a virtual camera evenly spaced around the unaligned 3D shape model with an elevation of 30 degrees. We only aggregate the multiple views during testing to re-
| Evaluation metric | NN | FT | ST | E | DCG | mAP |
|-------------------|----|----|----|---|-----|-----|
| Siamese [48]      | 0.405 | 0.403 | 0.548 | 0.287 | 0.607 | 0.469 |
| Shape2Vec [47]    | 0.620 | 0.628 | 0.684 | 0.354 | 0.741 | 0.650 |
| DCML [8]          | 0.650 | 0.634 | 0.719 | 0.348 | 0.766 | 0.674 |
| LWBR [52]         | 0.712 | 0.725 | 0.785 | 0.369 | 0.814 | 0.752 |
| DCA [3]           | 0.783 | 0.796 | 0.829 | 0.376 | 0.856 | 0.813 |
| SEM [40]          | 0.823 | 0.828 | 0.860 | 0.403 | 0.884 | 0.843 |
| DSSH [4]          | 0.831 | 0.844 | 0.886 | 0.411 | 0.893 | 0.858 |
| This paper        | 0.825 | **0.848** | **0.899** | 0.472 | 0.907 | **0.865** |

(a) SHREC13.

| Evaluation metric | NN | FT | ST | E | DCG | mAP |
|-------------------|----|----|----|---|-----|-----|
| Siamese [48]      | 0.239 | 0.212 | 0.316 | 0.140 | 0.496 | 0.228 |
| Shape2Vec [47]    | 0.714 | 0.697 | 0.748 | 0.360 | 0.811 | 0.720 |
| DCML [8]          | 0.272 | 0.275 | 0.345 | 0.171 | 0.498 | 0.286 |
| LWBR [52]         | 0.403 | 0.378 | 0.455 | 0.236 | 0.581 | 0.401 |
| DCA [3]           | 0.770 | 0.789 | 0.823 | 0.398 | 0.859 | 0.803 |
| SEM [40]          | **0.804** | 0.749 | 0.813 | 0.395 | 0.870 | 0.780 |
| DSSH [4]          | 0.796 | 0.813 | 0.851 | 0.412 | 0.881 | 0.826 |
| This paper        | 0.789 | **0.814** | **0.854** | 0.561 | 0.886 | **0.830** |

(b) SHREC14.

| Evaluation metric | NN | FT | ST | E | DCG | mAP |
|-------------------|----|----|----|---|-----|-----|
| Siamese [48]      | 0.118 | 0.076 | 0.132 | 0.073 | 0.400 | 0.067 |
| SEM [40]          | **0.840** | 0.634 | 0.745 | 0.526 | 0.848 | 0.676 |
| DSSH [4]          | 0.838 | 0.777 | 0.848 | 0.624 | 0.888 | 0.806 |
| This paper        | 0.816 | **0.799** | **0.891** | **0.685** | **0.910** | **0.831** |

(c) Part-SHREC14.

Table 4: **Comparison 3** to many-shot sketch-based 3D shape retrieval on SHREC13, SHREC14, and Part-SHREC14. Having a metric space revolving around semantic prototypes benefits five out of six metrics.

Results. Table 4 shows the results on all three benchmarks and six metrics. We compare to seven state-of-the-art baselines, which mostly focus on learning a joint feature space of sketches and 3D shapes with metric learning [6, 17, 43]. Across all three benchmarks, we observe the same trend, where we obtain the highest scores for five out of the six baselines. Only for the precision@1 metric (NN) do the recent approaches of Chen et al. [4] and Qi et al. [40] obtain higher scores on all three benchmarks. A reason for this behaviour is that both approaches directly optimize for the nearest neighbour metric. Indeed, Qi et al. [40] search in the label space while Chen et al. [4] perform a learned hashing. Overall, we conclude that our approach, while simple in nature, provides competitive results compared to the current state-of-the-art in sketch-based 3D shape retrieval.

Qualitative analysis. To gain insight in the pros and cons of our approach for retrieving 3D shapes from sketches, we provide qualitative examples in Figure 8. While rotations of unaligned shapes can be handled, confusion remains with visually and similar categories.

6. Conclusion

In this paper, we have introduced **open cross-domain visual search**. Rather than searching from one fixed source domain to one fixed target domain, the open setting strives to search among any number of domains. This translates into a search between any pair of source and target domains, a search from a combination of multiple sources, or a search within a combination of multiple targets. We have proposed a simple approach for open cross-domain search based on hyperspherical prototypes to align the semantics of multiple visual domains in a common space. Demonstrations on novel open cross-domain visual search tasks present how to search across multiple domains. State-of-the-art results on comparisons to existing closed cross-domain visual search tasks show the effectiveness of our approach.
Acknowledgements We thank Herke van Hoof for initial insight, Qing Liu for helpful correspondence, as well as Zenglin Shi and Hubert Banville for feedback. William Thong is partially supported by an NSERC scholarship.

References

[1] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In NeurIPS, 2011. 2
[2] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucco, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In CVPR, 2019. 2
[3] Jiaxin Chen and Yi Fang. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In ECCV, 2018. 8
[4] Jiaxin Chen, Jie Qin, Li Liu, Fan Zhu, Fumin Shen, Jin Xie, and Ling Shao. Deep sketch-shape hashing with segmented 3d stochastic viewing. In CVPR, 2019. 1, 2, 8
[5] Soumith Chintala, Marc’Aurelio Ranzato, Arthur Szlam, Yuandong Tian, Mark Tygert, and Wojciech Zaremba. Scale-invariant learning and convolutional networks. Applied and Computational Harmonic Analysis, 42(1):154–166, 2017. 2
[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In CVPR, 2005. 2, 8
[7] Gabriela Csurka. Domain adaptation in computer vision applications. Springer. 2
[8] Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Deep correlated metric learning for sketch-based 3d shape retrieval. In AAAI, 2017. 2, 8
[9] Jiankang Deng, Jia Guo, Nianman Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019. 2
[10] Soumak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In CVPR, 2019. 1, 2, 6
[11] Qi Dou, Daniel C Castro, Konstantinos Kamnitas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In NeurIPS, 2019. 2
[12] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In CVPR, 2019. 1, 2, 6
[13] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? ACM TOG, 31(4), 2012. 1, 2, 6, 7
[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. De-vice: A deep visual-semantic embedding model. In NeurIPS, 2013. 2, 3
[15] Yaroslav Ganin, Evgeniya Ustinova, Hanna Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. JMLR, 17(1), 2016. 2, 6
[16] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In NeurIPS, 2018. 2, 6
[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In CVPR, 2006. 2, 8
[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In NeurIPS-W, 2014. 3
[19] Conghui Hu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-classifier: sketch-based photo classifier generation. In CVPR, 2018. 2, 6, 7
[20] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil M Robertson, and Yongxin Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In ICCV, 2017. 2
[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, 2018. 4
[22] Bo Li, Yijuan Lu, Afzal Godil, Masaki Aono, Henry Johan, Jose M Saavedra, and Shoki Tashiro. Shrec13 track: large scale sketch-based 3d shape retrieval. In Eurographics workshop on 3D object retrieval, 2013. 1, 2, 6, 7
[23] Bo Li, Yijuan Lu, Afzal Godil, Tobias Schreck, Benjamin Bustos, Alfredo Ferreira, Takahiko Furuya, Manuel J Fonseca, Henry Johan, Takahiro Matsuda, et al. A comparison of methods for sketch-based 3d shape retrieval. CVIU, 119:57–80, 2014. 1, 2, 8
[24] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Hongbo Fu, Takahiko Furuya, Henry Johan, et al. Shrec14 track: Extended large scale sketch-based 3d shape retrieval. In Eurographics workshop on 3D object retrieval, 2014. 1, 2, 6, 7
[25] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In ICCV, 2017. 1
[26] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In CVPR, 2017. 6, 7
[27] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In ICCV, 2019. 2, 6
[28] Weiyang Liu, Yadong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In CVPR, 2017. 2
[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In ICLR, 2017. 4
[30] Peng Lu, Gao Huang, Yanwei Fu, Guodong Guo, and Hangyu Lin. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In CVPR, 2017. 6, 7
[31] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Tomas Mikolov. Distance-based sample. In NeurIPS, 2011. 2, 3
[32] Pascal Mettes, Elise van der Pol, and Cees GM Snoek. Hyperpherical prototype networks. In NeurIPS, 2019. 2
[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In ICLR, 2013. 3, 4
[34] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss metric learning using proxies. In ICCV, 2017. 2

[35] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In NeurIPS, 2009. 3

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In ICCV, 2019. 1, 2, 4

[37] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. arXiv:1710.06924, 2017. 2

[38] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In EMNLP, 2014. 3

[39] Bui Tuong Phong. Illumination for computer generated pictures. Communications of the ACM, 18(6):311–317, 1975. 7

[40] Anran Qi, Yi-Zhe Song, and Tao Xiang. Semantic embedding for sketch-based 3d shape retrieval. In BMVC, 2018. 2, 7, 8

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 115(3), 2015. 4

[42] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. ACM TOG, 2016. 1, 2, 6, 7

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, 2015. 2, 8

[44] Yuning Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In CVPR, 2018. 1, 2, 6

[45] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The Princeton shape benchmark. In Shape Modeling International, June 2004. 7

[46] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In ICML, 2013. 4

[47] Flora Ponjou Tasse and Neil Dodgson. Shape2vec: Semantic-based descriptors for 3d shapes, sketches and images. ACM TOG, 2016. 2, 8

[48] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In CVPR, 2015. 1, 2, 8

[49] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In CVPR, 2018. 2

[50] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In ECCV, 2016. 2

[51] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behave artistic media dataset for recognition beyond photography. In ICCV, 2017. 1

[52] Jin Xie, Guoxian Dai, Fan Zhu, and Yi Fang. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In CVPR, 2017. 1, 2, 8

[53] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In CVPR, 2018. 2

[54] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In ECCV, 2018. 1, 2, 6

[55] Andrew Zhai and Hao-Yu Wu. Making classification competitive for deep metric learning. In BMVC, 2019. 2

[56] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In CVPR, 2016. 6

[57] Junbao Zhuo, Shuhui Wang, Shuhao Cui, and Qingming Huang. Unsupervised open domain recognition by semantic discrepancy minimization. In CVPR, 2019. 2