UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Open Cross-Domain Visual Search

Thong, W.; Mettes, P.; Snoek, C.G.M.

DOI
10.1016/j.cviu.2020.103045
Publication date
2020
Document Version
Final published version
Published in
Computer Vision and Image Understanding

License
CC BY

Link to publication

Citation for published version (APA):
Thong, W., Mettes, P., & Snoek, C. G. M. (2020). Open Cross-Domain Visual Search. *Computer Vision and Image Understanding*, *200*, Article 103045. https://doi.org/10.1016/j.cviu.2020.103045

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:11 Oct 2026

Computer Vision and Image Understanding 200 (2020) 103045

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Open cross-domain visual search

William Thong *, Pascal Mettes, Cees G.M. Snoek

University of Amsterdam, Science Park 904, Amsterdam, The Netherlands

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

This paper addresses cross-domain visual search, where visual queries retrieve category samples from a different domain. For example, we may want to sketch an airplane and retrieve photographs of airplanes. Despite considerable progress, the search occurs in a closed setting between two pre-defined domains. In this paper, we make the step towards an open setting where multiple visual domains are available. This notably translates into a search between any pair of domains, from a combination of domains or within multiple domains. We introduce a simple – yet effective – approach. We formulate the search as a mapping from every visual domain to a common semantic space, where categories are represented by hyperspherical prototypes. Open cross-domain visual search is then performed by searching in the common semantic space, regardless of which domains are used as source or target. Domains are combined in the common space to search from or within multiple domains simultaneously. A separate training of every domain-specific mapping function enables an efficient scaling to any number of domains without affecting the search performance. We empirically illustrate our capability to perform open cross-domain visual search in three different scenarios. Our approach is competitive with respect to existing closed settings, where we obtain state-of-the-art results on several benchmarks for three sketch-based search tasks.

1. Introduction

This paper aims for visual category search across domains. The task is to retrieve visual examples from a specific category in one domain, given a query from another domain. For example, we may want to retrieve *images* of an "airplane" from a quickly-drawn *sketch*. Cross-domain visual search has made considerable progress, showing the possibility to retrieve natural images (Eitz et al., 2010; Sangkloy et al., 2016) or 3D shapes (Li et al., 2013, 2014b,a) from sketches. Different from existing works, which emphasize retrieval from a single source domain to a single target domain, we open the search beyond two domains. The motivation for a search among many domains is that in practice, categories come in many forms (Peng et al., 2019; Wilber et al., 2017; Li et al., 2017). Hence, we may have queries from several source domains, or want to search with any possible combination of source and target domains. For example, we may now want to combine a *sketch* and a *clipart* of an "airplane" to retrieve *photograph* samples, or use a *clipart* of an "airplane" to retrieve *3D shapes*. In this paper, we strive for such an open setting: we visually search for categories from any source domain to any target domain, with the ability to search from and within multiple domains simultaneously.

Within cross-domain visual search, an important challenge is the gap between source and target domains (Shen et al., 2018; Yelamarthi et al., 2018; Dey et al., 2019; Dutta and Akata, 2019; Xie et al., 2017; Chen et al., 2019). Given the inherent difference in representations, reducing the domain gap is an intuitive solution. Both Shen et al. (2018) and Yelamarthi et al. (2018) have highlighted the importance of domain adaptation losses for cross-domain search, especially when searching for unseen categories. Yet, relying on domain adaptation methods makes the search unsuited for an open setting by design, due to the requirement of pair-wise domain training. As a consequence, opening the search to many domains creates new challenges as (*i*) all domains should to be mapped to a unique embedding space, and (*ii*) new domains should be able to be added continuously in an efficient fashion. We address the challenges of open cross-domain visual search.

Inspired by recent works on prototype-based embedding spaces (Movshovitz-Attias et al., 2017; Wen et al., 2016; Snell et al., 2017), we introduce prototype learners for cross-domain visual search in an open setting. Prototype learning has shown to simplify model training and improve performance for image retrieval (Movshovitz-Attias et al., 2017; Wen et al., 2016) and classification (Snell et al., 2017) problems in a low-shot setting. In this work, we leverage prototype learners to perform visual search across multiple domains simultaneously. We define prototypes to unite all domains. Inputs from every domain are mapped to a common semantic space, where every learner is domain-specific and is trained separately. During training, the semantic space is defined by categorical prototypes, corresponding to word embeddings of category names. Learning then consists of regressing inputs to their corresponding categorical prototype in this common semantic space,

* Corresponding author.
E-mail address: w.e.thong@uva.nl (W. Thong).

https://doi.org/10.1016/j.cviu.2020.103045
Received 1 April 2020; Received in revised form 17 July 2020; Accepted 21 July 2020
Available online 29 July 2020
1077-3142/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

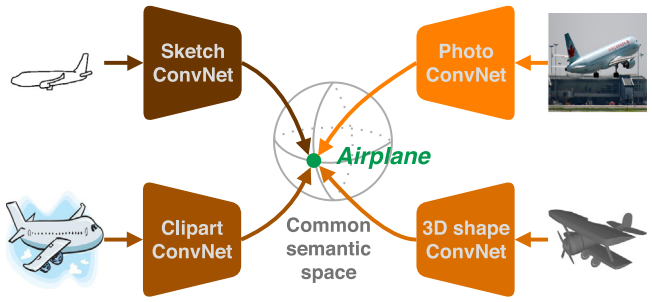


Fig. 1. Open cross-domain visual search. We search for categories from any number of source domains to any number of target domains. Mapping examples to a common semantic space enables any possible combinations of domains when searching for categories.

as illustrated in Fig. 1. Query representations for search are further refined with neighbors 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 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 retraining previous models.

Empirically, we first demonstrate the ability to perform open cross-domain visual search, highlighting new applications and search possibilities, *i.e.* (*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. Second, while designed for the open cross-domain setting, our approach also works in the conventional closed settings, allowing for comparisons to current approaches. We compare to sketch-based image and 3D shape retrieval, usually considered separately in the literature. We show the versatility of our approach to handle them. Across three well-established tasks totalling seven benchmarks, we obtain state-of-the-art results, which highlights the effectiveness of focusing solely on the semantic space for cross-domain search.

Contributions. Our main contribution is the introduction of open cross-domain visual search. We open the search to many domains, with the ability to retrieve categories from and among any number of domains. To achieve this, we introduce a simple prototype learner for each domain to learn a common semantic space efficiently. Empirically, solely relying on semantic prototypes turns into an effective solution for cross-domain visual search in both newly proposed open settings and existing closed settings. All code and setups are released to foster further research in open cross-domain visual search.[1]

2. Related work

We first cover related work in cross-domain search, where a large body of works focuses on retrieving natural images or 3D shapes from sketches. We then review relevant work addressing multiple domains and on how to learn semantic spaces with prototype learners.

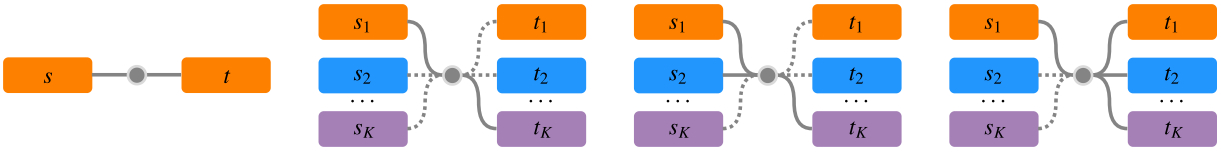
Cross-domain image search. Sketch-based image retrieval has been a topic of vision community interest for a long time (Kato, 1992; Jacobs et al., 1995). The seminal work of Eitz et al. (2010) established the first benchmark for its evaluation, which led to the construction of common descriptors for sketches and images, such as bag-of-features (Eitz et al., 2010), bag-of-regions (Hu et al., 2011), histogram of oriented gradients (Hu and Collomosse, 2013), or specialized descriptors for edges (Saavedra, 2014). With the resurgence of convolutional networks, the dominant approach has shifted towards the learning of a joint semantic space of sketches and images. Qi et al. (2016) learn a joint embedding with a Siamese network while Bui et al. (2017) rely on a triplet network. Bui et al. (2018) add a classification head with a multi-stage training to make features even more discriminative. In all these works, the semantic spaces model categories implicitly, as they rely on sample-based methods such as the Siamese (Hadsell et al., 2006; Chopra et al., 2005) or triplet (Schroff et al., 2015; Weinberger and Saul, 2009) losses to learn cross-domain visual similarities. In this paper, we explicitly define semantic representations for every category in the embedding space. This removes the need for sampling and mining of cross-domain pairs, resulting in a simpler training procedure.

Sketch-based image retrieval is also considered as a zero-shot learning problem (Shen et al., 2018; Yelamarthi et al., 2018). In this context, a common approach is to bridge the domain gap between sketches and images. Shen et al. (2018) fuse sketch and image representations with a Kronecker product, while Yelamarthi et al. (2018) introduce domain confusion with generative models to produce domain-agnostic features. Dey et al. (2019) combine gradient reversal layers with metric learning losses to extract the mutual information from both domains. Dutta and Akata (2019) tie the semantic space with visual features from both domains by learning to generate them while Dutta and Biswas (2019) prefer to separate them explicitly. Alternatively, Liu et al. (2019) preserve the knowledge from a pre-trained model to avoid features to drift away during training. Hu et al. (2018a) have also explored how to synthesize classifiers derived from sketches for few-shot image classification. By focusing on domain adaptation, current approaches are optimized to map from a single specific source domain to a single specific target domain. Instead, we consider cross-modal image search from any number of source domains to any number of target domains.

Cross-domain 3D shape search. Searching for 3D shapes from a sketch has been accelerated by the SHREC challenges (Li et al., 2013, 2014b,a). A common approach is to transform the 3D shape search into an image search problem by projecting the unaligned 3D shape into multiple 2D views (Su et al., 2015). In this regard, the main methodological approach is to learn a joint embedding space of sketches and 2D view renderings of the unaligned 3D shapes. Wang et al. (2015) map both sketches and shapes in a similar feature space with a Siamese network, while Tasse and Dodgson (2016) learn to regress to a semantic space with a ranking loss. Dai et al. (2017) correlate both sketch and 3D shape representations to bridge the domain gap. Xie et al. (2017) employ the Wasserstein distance to create a barycentric representation of shapes. Qi et al. (2018) apply loss functions on the probabilistic label space rather than the feature space. Chen et al. (2019) propose an advanced sampling of 2D views for the unaligned shapes. Learning cross-domain visual similarities with Siamese or triplet losses typically requires a multi-stage training or negative sampling schemes. A prototype learner removes this requirement, and enables the addition of new domains without the need for retraining existing models.

Searching beyond two domains. Using multiple domains has been investigated in unsupervised domain adaptation (Peng et al., 2017; Csurka, 2017) and unsupervised domain generalization (Blanchard et al., 2011), where the task is to classify unlabeled target samples by learning a classifier on labeled source samples. As such, Peng et al. (2019) illustrate how challenging classification becomes when multiple domains are considered. A new challenge then arises as classifiers have to be designed to benefit from the inherent gap among multiple domains (Xu et al., 2018; Peng et al., 2019; Zhuo et al., 2019; Dou et al., 2019; Carlucci et al., 2019). In this paper, we focus on a different multi-domain task: we consider cross-domain retrieval where category labels are present for both source and target domains, and where the main challenge is to learn a common embedding space for all domains.

[1] Source code is available at https://github.com/twuilliam/open-search.



(a) *one* source to *one* target (b) *any* source to *any* target (c) *many* sources to *any* target (d) *any* source to *many* targets

Fig. 2. Cross-domain image search focuses on mapping (a) from one fixed source to one fixed target domain. In this paper, we consider an open domain setting with K available domains. We search (b) from any source to any target domain, (c) from multiple source domains to any target domain, and (d) from any source domain to multiple target domains.

Prototype learners. Learning metric spaces with prototypes for image retrieval (Sohn, 2016; Wen et al., 2016; Movshovitz-Attias et al., 2017; Zhai and Wu, 2019; Liu et al., 2017b; Wang et al., 2018; Deng et al., 2019; Snell et al., 2017) and classification (Mensink et al., 2013; Snell et al., 2017; Chintala et al., 2017; Mettes et al., 2019) provides a simpler alternative to common contrastive (Hadsell et al., 2006; Chopra et al., 2005) or triplet (Schroff et al., 2015; Weinberger and Saul, 2009) loss functions. One line of work learns to regress to moving prototypical representations. Depending on the task, such prototypes can correspond to center (Wen et al., 2016), proxy (Movshovitz-Attias et al., 2017; Zhai and Wu, 2019), or support (Snell et al., 2017; Ren et al., 2018) representations. While the distance measure usually relies on a cosine or Euclidean distance, a margin has also been introduced in the distance measure (Liu et al., 2017b; Wang et al., 2018; Deng et al., 2019). Another line of work regresses to fixed prototypical representations to avoid the simultaneous learning of prototypes and model parameters. Examples of fixed representations include class means (Mensink et al., 2013), one-hot representations (Chintala et al., 2017), or separated representations (Mettes et al., 2019). We build on the latter approach for open cross-domain visual search. We formulate semantic prototypes to align examples from many domains simultaneously. Categories are represented by fixed semantic prototypes in the embedding space. We then define a prototype learner for every domain to map visual inputs to the common space where open cross-domain search occurs.

3. Method

3.1. Problem formulation

Fig. 2 illustrates the search scenarios for open cross-domain search. While the *closed* cross-domain setting focuses on one pre-defined source s and one pre-defined 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 visual search.

Formally, let $\mathcal{D}$ 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 $\mathcal{T} = \{(\mathbf{x}_n^d, y_n)\}_{n=1}^N$, where $\mathbf{x}_n^d \in \mathcal{I}_d$ denotes an input example from a visual domain $d \in \mathcal{D}$ 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 $\mathbf{x}^{d_i}$ 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 a pre-defined source domain and a pre-defined 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 $(\mathbf{x}^{d_i}, \mathbf{x}^{d_j})$ of two different domains $i \neq j$ to search in the gallery of a third domain k. Conversely, given a sample query $\mathbf{x}^{d_i}$, we can search in the combined gallery of multiple target 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. Third, we outline how open cross-domain search occurs.

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 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 (Frome et al., 2013; Palatucci et al., 2009). In this work, we opt for word embeddings, *e.g.*, word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), 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 is formulated as a convolutional network followed by an ℓ_2-normalization on the D-dimensional network outputs.

We propose the following function to map an example $\mathbf{x}^d$ of domain d to its categorical prototype $\phi(y)$ in the common semantic space:

$$p(y|\mathbf{x}^d, d) = \frac{\exp\Big(-s \cdot c\big(f_d(\mathbf{x}^d), \phi(y)\big)\Big)}{\sum_{y' \in \mathcal{Y}} \exp\Big(-s \cdot c\big(f_d(\mathbf{x}^d), \phi(y')\big)\Big)}, \tag{1}$$

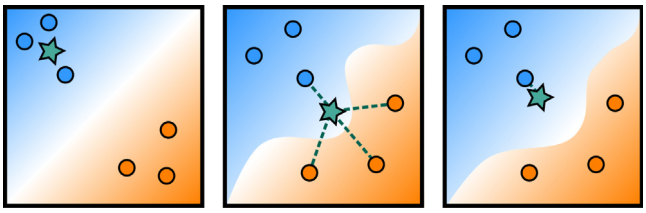
where $s \in \mathbb{R}_{>0}$ denotes a scaling factor, inversely equivalent to the temperature (Hinton et al., 2014). Intuitively, the scaling controls how samples are spread around categorical prototypes. $c(\cdot, \cdot)$ is defined as the cosine distance:

$$c(f_d(\mathbf{x}^d), \phi(y)) = 1 - \langle f_d(\mathbf{x}^d), \phi(y) \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the dot product. As both $f_d(\mathbf{x})$ and $\phi(y)$ lie on the hypersphere $\mathbb{S}^{D-1}$, they have a unit norm. Finally, learning every mapping function f_d is done by minimizing the cross-entropy loss over the training set:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \log p(y_n|\mathbf{x}_n^d, d). \tag{3}$$

In our approach, the representations of the categorical prototypes remain unaltered. Hence, we only take the partial derivative with respect to the mapping function parameters. When training the mapping function f_d for domain d, only examples $\mathbf{x}^d$ of domain d are used as inputs.



(a) Ideal. (b) Real. (c) Refined.

Fig. 3. Cross-domain query refinement. (a) Ideally, the neighborhood of the query (star) is only close to examples from the same category. (b) In reality, variability causes noise in the semantic space. Hence, the query might also be close to samples from other categories. (c) We tackle this variability by refining the query representation.

Searching across open domains. In the search evaluation phase, similarity between source and target samples is measured with the cosine distance in the shared semantic space. 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 source query. As all domains map to the same common semantic space, domains can straightforwardly be combined either to search with queries from multiple domains or to search within a gallery of multiple domains.

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 then propose to refine the initial query representation using a nearby example from the target domain, as illustrated in Fig. 3.

We refine the query representation p_0 by performing a spherical linear interpolation with a relevant representation p_1. The refined representation $\hat{p}$ is:

$$\hat{p}(p_0, p_1|\lambda) = \frac{\sin\big((1-\lambda)\Omega\big)}{\sin\Omega} p_0 + \frac{\sin\big(\lambda\Omega\big)}{\sin\Omega} p_1, \tag{4}$$

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. In retrieval, we set p_1 as the 1-nearest neighbor of p_0 in the target set. This mixture does not require any label and relies on the fact that the recall at one is usually very high. In classification, p_1 is the word embedding of the category name.

4. Open cross-domain visual search

In the first set of experiments, we demonstrate the 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 without hassle. Second, we show the potential and positive effect of searching from multiple source domains for any target domain. Third, we exhibit the possibility of searching in multiple target domains simultaneously.

Setup. We evaluate on the recently introduced *DomainNet* (Peng et al., 2019), which contains 596,006 images from 345 classes. Images are gathered from six visual domains: *clipart*, *infograph*, *painting*, *pencil*, *photo* and *sketch*. We consider retrieval in *zero-* and *many*-shot evaluations: (*i*) in the *zero*-shot evaluation, $\mathcal{Y}$ is split into $\mathcal{Y}_{train}$ and $\mathcal{Y}_{test}$, with $\mathcal{Y}_{train} \cap \mathcal{Y}_{test} = \emptyset$, *i.e.*, categories to be searched during inference have not been seen during training; (*ii*) the *many*-shot evaluation uses the same categories during both training and testing. The zero-shot evaluation randomly splits samples into 300 training and 45 testing classes. Following the zero-shot learning good practices in Xian et al. (2018), we have verified the presence of the 345 categories of *DomainNet* (Peng et al., 2019) in *ImageNet* (Russakovsky et al., 2015), where we identify 188 separate categories. From this list of separate categories, we randomly sample 45 zero-shot categories with at least 40 samples per class in every domain. The many-shot evaluation follows the original splits from Peng et al. (2019). We report the mean average precision (mAP@all).

Implementation details. Throughout the paper and unless stated otherwise, we use SE-ResNet50 (Hu et al., 2018b) pre-trained on ImageNet (Russakovsky et al., 2015) as a backbone, and word2vec trained on a Google News corpus (Mikolov et al., 2013) as the common semantic space. We remove the final classifier layer of SE-ResNet50, and replace it with a fully-connected layer of size $D = 300$ initialized with random weights. The new layer is followed by a linear activation and batch normalization (Ioffe and Szegedy, 2015). We optimize the loss in Eq. (3) with Nesterov momentum (Sutskever et al., 2013) by setting the coefficient to 0.9. We apply a learning rate of 1e−4 with cosine annealing without warm restarts (Loshchilov and Hutter, 2017) and a batch size of 128. We use a scaling factor s of 20, and decrease it to 10 for Sections Sections 5.2 and 5.3. We set $\lambda = 0.7$ when evaluating on unseen classes (*i.e.*zero-shot and few-shot evaluations) and to 0.4 when evaluating on seen classes (*i.e.*many-shot evaluation). The implementation rests on the Pytorch (Paszke et al., 2019) framework and image similarities are computed with the Faiss (Johnson et al., 2017) library. Word embeddings of class names are extracted with the Gensim (Řehůřek and Sojka, 2010) library.

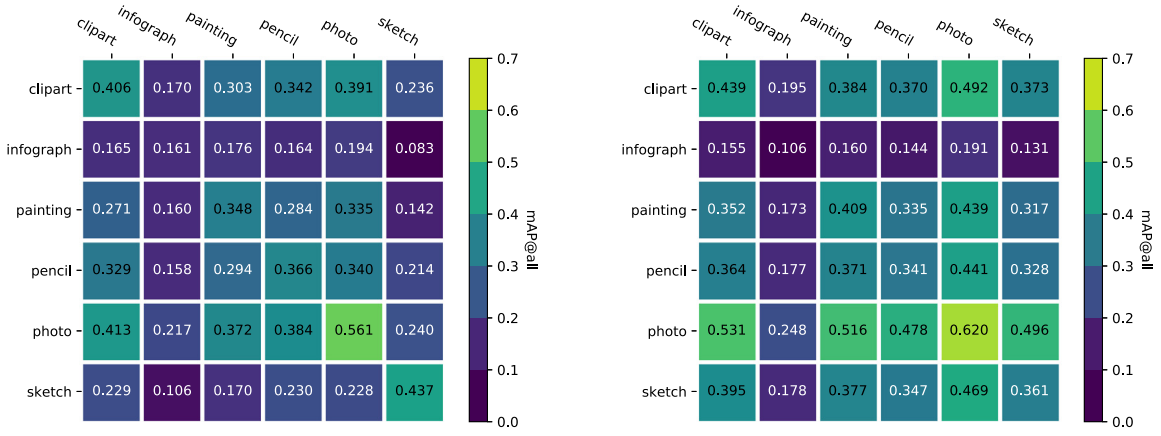
4.1. From any source to any target domain

First, we demonstrate how searching from any source to any target domain in an open setting is trivially enabled by our approach. Fig. 4 shows the result of 72 cross-domain search evaluations; corresponding to all six cross-domain pairs for both zero- and many-shot evaluations. 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 methods – 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.

In the zero-shot evaluation with an evaluation on the unseen classes (Fig. 4a), 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.

In the many-shot evaluation with an evaluation on all classes (Fig. 4b), the *photograph* domain exhibits a similar behavior. Though, in this case the search performance for *sketches* is at the same level as other considered domains, such as *clipart*, *painting* or *pencil*. Seeing all classes helps the prototype learner to better grasp the variability in *sketches*. The *infograph* domain remains the most challenging. 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 for both zero- and many-shot evaluations since we bypass the need to align different domains.

We quantitatively compare with the state-of-the-art SAKE (Liu et al., 2019) on zero-shot sketch-based image retrieval. We run SAKE from the original source code provided by the authors. Table 1 presents the results when considering sketches as the source domain and retrieving

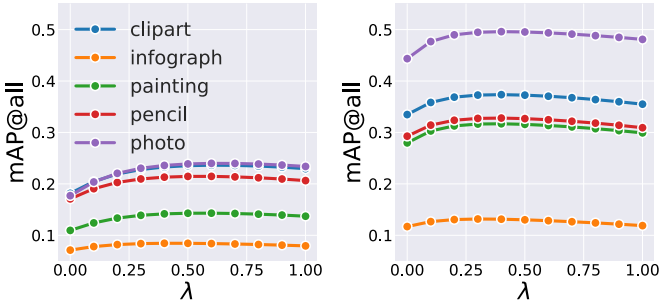


(a) *Zero*-shot evaluation (on 45 unseen classes). (b) *Many*-shot evaluation (on all 345 classes).

Fig. 4. Demonstration 1 for visual search from any source (columns) to any target (rows) domain in mAP@all. Our approach can perform 36 cross-domain searches for both (a) *zero*-shot evaluation, and (b) *many*-shot evaluation, without any modifications as we bypass the need to align domains.

Table 1
Visual search from sketches as a source to any target domain comparison with SAKE (Liu et al., 2019) in mAP@all. Our formulation achieves competitive results in both *zero*- and *many*-shot evaluations.

Target domain	*zero*-shot		*many*-shot	
	SAKE	*This paper*	SAKE	*This paper*
clipart	0.199	**0.236**	0.268	**0.373**
infograph	0.080	**0.083**	0.097	**0.131**
painting	0.118	**0.142**	0.203	**0.317**
pencil	0.181	**0.214**	0.230	**0.328**
photo	0.206	**0.240**	0.358	**0.496**



(a) *Zero*-shot evaluation (b) *Many*-shot evaluation

Fig. 5. Ablation on cross-domain query refinement on DomainNet, with *sketches* as a source. Refining the source representation always improves the retrieval performance.

Table 2
Demonstration 2 for visual search from multiple sources to any target domain (absolute improvement in mAP@all). In our approach, searching from multiple sources is as easy as using a single source, as we only have to average their positions in the common semantic space. Searching (a) from multiple diverse domains is preferred when the source is less informative, while (b) more examples from the same domain are preferred when the source is more informative.

(a) Improving the less informative sketch representations

Target domain	*zero*-shot			*many*-shot		
	sk+sk	sk+in	sk+ph	sk+sk	sk+in	sk+ph
clipart	+.057	+.072	**+.211**	+.097	+.036	**+.178**
infograph	+.018	+.067	**+.107**	+.031	+.002	**+.075**
painting	+.035	+.080	**+.186**	+.079	+.029	**+.154**
pencil	+.054	+.060	**+.154**	+.083	+.043	**+.156**
photo	+.064	+.112	**+.328**	+.127	+.049	**+.185**

(b) Improving the more informative photograph representations

Target domain	*zero*-shot			*many*-shot		
	ph+ph	ph+in	ph+sk	ph+ph	ph+in	ph+sk
clipart	**+.070**	+.012	+.048	**+.075**	+.002	+.067
infograph	**+.029**	−.035	+.005	**+.027**	−.062	+.018
painting	**+.052**	+.011	+.008	**+.061**	+.004	+.049
pencil	**+.054**	+.012	+.037	**+.066**	+.000	+.057
sketch	+.041	+.001	**+.202**	**+.075**	−.013	−.030

images in any of the other domains. SAKE has been proposed with a zero-shot evaluation design from the start, which makes it strong in this setting. Indeed, results are close, we only observe an improvement of 0.3% (*infograph*) up to 3.7% (*clipart*). When the evaluation focuses on a large number of categories, we notice higher gains from 3.4% (*infograph*) up to 13.8% (*photograph*) in the many-shot evaluation. Our embedding space is better partitioned for all categories thanks to the semantic prototypes. Overall, our formulation provides competitive performance in both zero- and many-shot evaluations with a simpler training procedure.

Finally, we also assess the importance of the proposed refinement module of Eq. (4). Fig. 5 illustrates the effect of our cross-domain prototypical refinement when searching in any target domain from the *sketch* domain. We create a mixture between the *sketch* query ($\lambda = 0$) and its nearest neighbor in the gallery ($\lambda = 1$) for retrieval. For both zero- and many-shot evaluations, refining the representations improves the performance. We observe a need for a lower mixture for the many-shot evaluation, as classes are all seen during training compared with the zero-shot evaluation. Refining the representations helps to bridge the inherent cross-domain gap.

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. 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 2a 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 evaluation, including *infographs*

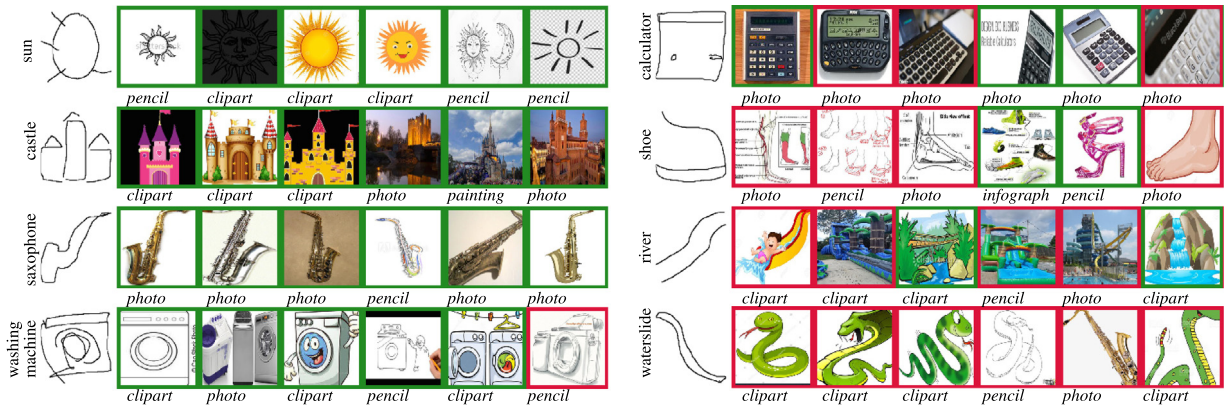


Fig. 6. Demonstration 3 for visual search from any source to multiple target domains. Correct results are in *green*, incorrect results in *red*. For abstract categories such as "sun", abstract domains such as *clipart* or *pencil* drawings tend to be retrieved first. When *sketches* are more ambiguous such as "calculator", some retrieved results are incorrect but resemble the shape. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and *photographs* improves upon sketch-based search only. In the many-shot evaluation, including *infographs* improves upon search by one *sketch*, but not by two *sketches*, which is not surprising given the low scores for infographs individually. *Photographs* with *sketches* obtain the highest scores, regardless of the target domain or the evaluation setting.

We also consider a more challenging multi-source search scenario where we search from the most informative source (*photograph*) and one of the least informative sources (*infograph* or *sketch*). Table 2b confirms the positive effect of searching with an additional domain. Adding *infographs* only improves the results marginally. Performance can even decrease when searching within one of the least informative domains, because the combination creates a destructive noise that moves the initial representation to a wrong direction. Adding *sketches* can benefit searching within *sketches* when the uncertainty is high, as in a zero-shot evaluation, but slightly decreases the score when the uncertainty is low, as in a many-shot evaluation. In the other target domains, *sketches* are much more effective than *infographs* when added to *photographs*. Though, the improvement is lower than searching from two *photographs*. When searching from an informative source domain, combining it with itself improves more than a combination with a less informative domain for both zero- and many-shot evaluations.

This demonstration shows the potential of searching from multiple sources. It is better to diversify the search by using multiple diverse domains when the source is less informative while more queries from the same domain are preferred when the source is more informative. 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 in a many-shot evaluation.

Fig. 6 provides qualitative results for eight *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 categories.

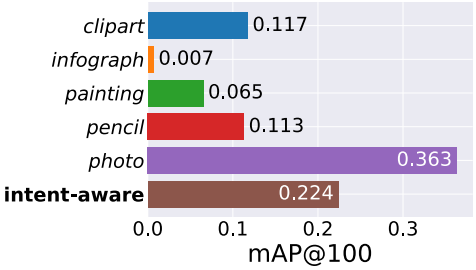


Fig. 7. Intent-aware evaluation for visual search from sketches to the other five target domains. Correct retrieved images in the top-ranked results more likely come from the *photograph* than the *infograph* domain.

We also quantitatively measure the retrieval performance when searching from sketches to the other five target domains simultaneously. When computing the mAP@100, we obtain a score of 0.565. Though, this measure does not take into account the differences and diversity among domains, as it considers all of them as similar. As such, we report the intent-aware mAP (Agrawal et al., 2009). Extending the mAP to an intent-aware formulation provides an estimate of the result diversity by: (i) computing the mAP per domain, and (ii) summing them with a weighting that corresponds to the occurrences of every category within each domain. Fig. 7 shows the per domain and intent-aware mAP@100. The *photograph*-mAP@100 is the highest score, which indicates correct *photographs* are in the top-ranked results compared with other target domains. The *infograph*-mAP@100 obtains the lowest score, which means that there are very few correct *infographs* in the top-ranked results. When the differences among domains are taken into consideration, the intent-aware mAP@100 results in 0.224. In a search within multiple domains, the informativeness of each domain influences the top-ranked results.

5. Closed cross-domain visual search

Our approach is geared towards open cross-domain visual search, as demonstrated in the previous section. To get insight in the effectiveness

Table 3
Comparison 1 to zero-shot sketch-based image retrieval on TU-Berlin Extended and Sketchy Extended. Aligning solely the semantics improves cross-domain image retrieval.

(a) Real-valued representations

	TU-Berlin extended		Sketchy extended	
	mAP@all	prec@100	mAP@all	prec@100
EMS (Lu et al., 2018)	0.259	0.369	n/a	n/a
CAAE (Yelamarthi et al., 2018)	n/a	n/a	0.196	0.284
ADS (Dey et al., 2019)	0.110	n/a	0.369	n/a
SEM-PCYC (Dutta and Akata, 2019)	0.297	0.426	0.349	0.463
SG (Dutta and Biswas, 2019)	0.254	0.355	0.376	0.484
SAKE (Liu et al., 2019)	0.475	**0.599**	0.547	0.692
This paper	**0.517**	0.557	**0.649**	**0.708**

(b) Binary representations

	TU-Berlin extended		Sketchy extended	
	mAP@all	prec@100	mAP@all	prec@100
EMS (Lu et al., 2018)	0.165	0.252	n/a	n/a
ZSIH (Shen et al., 2018)	0.220	0.291	0.254	0.340
SEM-PCYC (Dutta and Akata, 2019)	0.293	0.392	0.344	0.399
SAKE (Liu et al., 2019)	0.359	0.481	0.364	0.487
This paper	**0.404**	**0.517**	**0.466**	**0.618**

(c) Generalized setting

	TU-Berlin extended		Sketchy extended	
	mAP@all	prec@100	mAP@all	prec@100
ZSIH (Shen et al., 2018)	0.142	0.218	0.219	0.296
SEM-PCYC (Dutta and Akata, 2019)	0.192	**0.298**	0.307	0.364
SG (Dutta and Biswas, 2019)	0.149	0.226	0.331	0.381
This paper	**0.211**	0.224	**0.397**	**0.421**

of our approach for cross-domain visual search in general, we also perform an extensive comparative evaluation on standard cross-domain settings, which search between two domains. In total, we compare on three of the most popular cross-domain search tasks, namely zero-shot sketch-based image retrieval (Sangkloy et al., 2016; Shen et al., 2018), few-shot sketch-based image classification (Hu et al., 2018a), and many-shot sketch-based 3D shape retrieval (Li et al., 2013, 2014b). 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. Below, we present each comparison separately.

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. *TU-Berlin Extended* (Eitz et al., 2012; Zhang et al., 2016) contains 20,000 sketches and 204,070 images from 250 classes. Following Shen et al. (2018), we select 220 classes for training and 30 classes for testing. *Sketchy Extended* (Sangkloy et al., 2016; Liu et al., 2017a) contains 75,481 sketches and 73,002 images from 125 classes. Similarly, following Shen et al. (2018), we select 100 classes for training and 25 classes for testing. For fair comparison with Liu et al. (2019), we select the same unseen classes for both datasets. Following recent works (Shen et al., 2018; Dutta and Akata, 2019; Liu et al., 2019), we report the mAP@all and the precision at 100 (prec@100) scores.

Results. Table 3a compares to six 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 (Ganin et al., 2016; Gonzalez-Garcia et al., 2018). On Sketchy Extended, our approach outperforms other baselines. On TU-Berlin Extended, we obtain the highest mAP@all score, while the recently introduced SAKE by Liu et al. (2019) obtains a higher prec@100 score. SAKE is better at grouping images from the same category together thanks to the preservation module that produces tightly distributed representations. Our method is better at retrieving relevant images in the first ranks as the refinement module reduces the noise in the query representations.

Following previous work in zero-shot sketch-based image retrieval (Lu et al., 2018; Shen et al., 2018; Dutta and Akata, 2019; Liu et al., 2019), we also report the retrieval performance on binary representations. As previously proposed in Dutta and Akata (2019) and Liu et al. (2019), real-valued representations are projected to a low-dimensional space and quantized with iterative quantization (Gong et al., 2012). We compute the transformation on the training set and apply it on both sketch and image testing sets. Note that we first refine the representations, then apply iterative quantization. Table 3b compares the proposed formulation with binary representations of 64 dimensions. Compared with real-valued representations in Table 3a, we notice a higher drop in the mAP@all score than the prec@100 score. Compared with other baselines, our semantic space based on word embeddings better preserves the information when compressed to a low-dimensional space.

As recently introduced by Dutta and Akata (2019), we also evaluate on a generalized setting in Table 3c, where the gallery set also includes images from seen classes. Following their protocol, we reserve 20% of the samples from the seen classes for evaluation and use VGG16 (Simonyan and Zisserman, 2014) in this experiment for fair comparison. On Sketchy Extended, our approach also outperforms other baselines. On TU-Berlin Extended, we obtain the highest mAP@all score, while SEM-PCYC by Dutta and Akata (2019) obtains a higher prec@100 score. Similar to the zero-shot evaluation, our method is better at ranking images than grouping them together. Overall, focusing solely on semantic alignment outperforms alternatives on domain adaption or knowledge preservation across three different settings derived from two datasets.

To understand the effect of the distance scaling hyper-parameter defined in Eq. (1), we vary its value on both datasets in Fig. 8. We observe the same behavior on both datasets. When $s = 1$ as in a common softmax function, it yields the lowest results. A higher scaling helps to narrow the probability distribution, resulting in a better retrieval performance. There is a tipping point around $s = 20$, after which performance decreases. Calibrating the softmax with a high distance scaling factor improves the retrieval performance.

Table 4
Comparison 2 to few-shot sketch-based image classification on a subsampled Sketchy Extended (multi-class accuracy). Our metric learning approach outperforms model regression approaches.

	w2v	Sketch		Image	
		one-shot	*five*-shot	*one*-shot	*five*-shot
M2M (Hu et al., 2018a)	n/a	n/a	79.93	n/a	93.55
F2M (Hu et al., 2018a)	35.90	68.16	83.01	84.12	93.89
This paper	**80.39**	**82.19**	**85.13**	**90.63**	**94.63**

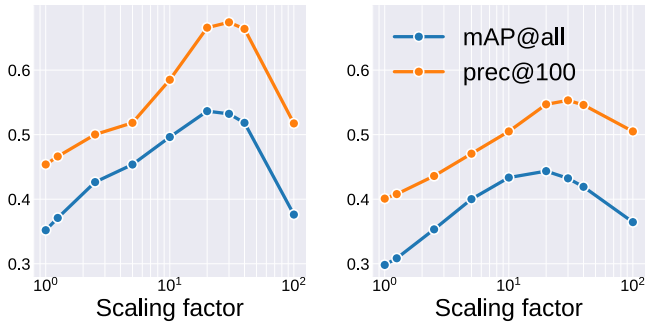
Qualitative analysis. To understand which sketches trigger the performance of natural image retrieval, we provide several qualitative sketch queries with their top retrieved images in Fig. 9. Our approach works well for typical sketches of categories. For example, the "cup" or "parrot" sketches exhibit a typical definition of their respective categories. In return, the search is very effective despite the variation in image appearance and viewpoints. Results degrade when sketches are ambiguous or in non-canonical views. For example, the "tree" sketch can easily be confused with the smoke ring of a "volcano" or the shape of a "windmill". Typical shape drawings of sketches matter for zero-shot image retrieval.

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 have not been observed during training. Different from the zero-shot retrieval scenario, the few-shot classification evaluation 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 (Sangkloy et al., 2016; Liu et al., 2017a). For fair comparison with Hu et al. (2018a), we subsample the Sketchy Extended to match the size of their private split. We select the same 115 classes for training and 10 classes for testing. We also rely on VGG19 (Simonyan and Zisserman, 2014) as a backbone. We evaluate the performance with the multi-class accuracy. Classification is done by measuring the distance to the class prototypes. Following Hu et al. (2018a), we evaluate on three different modes by setting the prototypes of the unseen classes to: (i) word vectors (w2v), (ii) *one* or *five* sketch representations, and (iii) *one* or *five* image representations. The latter is considered as an upper-bound of this cross-domain task. Following Hu et al. (2018a), the model is trained once and we report the average classification accuracy over 500 runs with different sets of sketches or images in the few-shot evaluation.

Results. Table 4 compares our formulation to two baselines introduced by Hu et al. (2018a). 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, the biggest relative improvement is achieved in the one-shot evaluation. It is also interesting to compare the w2v and one-shot sketch evaluation modes. As the one-shot sketch exhibits a higher score, it means that sketch representations capture visual details that cannot be described with word representations only. Our approach is also effective for cross-domain classification, especially with low shots.

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 Fig. 10. 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. For example, Fig. 10a shows a well sketched "cat" in one of the canonical positions while Fig. 10b exhibits a "cat" without any whiskers and in a strange view as we only see the face. Another important assertions is the sketch separability. For example, the "airplane" sketch in Fig. 10b could be confused with a "knife". Appearance, viewpoint and separability matter when relying on sketches for few-shot image classification.



(a) Sketchy Extended (b) TU-Berlin Extended

Fig. 8. Scaling hyper-parameter ablation. We evaluate the scaling of the softmax function. $s = 20$ yields the best results for both datasets, especially for the mAP@all score.

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* (Li et al., 2013) is constructed from the TU-Berlin (Eitz et al., 2012) and Princeton Shape Benchmark (Shilane et al., 2004) datasets, resulting in 7200 sketches and 1258 3D shapes from 90 classes. The training set contains 50 sketches per class, the testing set 30. *SHREC14* (Li et al., 2014b) contains more 3D shapes and more classes, resulting in 13,680 sketches and 8987 3D shapes from 171 classes. The training and testing splits of sketches follow the same protocol as SHREC13. We also report on *Part-SHREC14* (Qi et al., 2018), which contains 3840 sketches and 7238 3D shapes from 48 classes. The sketch splits also follow the same protocol, while the 3D shapes are now split into 5812 for training and 1426 for testing to avoid overlap.

Following previous works (Chen and Fang, 2018; Xie et al., 2017; Su et al., 2015), we generate 2D projections for all 3D shape models using the Phong reflection model (Phong, 1975). Similarly, we 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 reduce complexity. We report six retrieval metrics (Li et al., 2014a). The nearest neighbor (NN) denotes precision@1. The first tier (FT) is the recall@K, where K is the number of 3D shape models in the gallery set of the same class as the query. The second tier (ST) is the recall@$2K$. The E-measure (E) is the harmonic mean between the precision@32 and the recall@32. The discounted cumulated gain (DCG) and mAP are also reported.

Results. Table 5 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 (Hadsell et al., 2006; Chopra et al., 2005; Schroff et al., 2015). 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. (2019) and Qi et al. (2018) obtain higher scores on all three benchmarks. A first reason for this behavior is that both approaches directly optimize for the nearest neighbor metric. Qi et al. (2018) search in the label space while Chen et al. (2019) perform a learned hashing. A second reason comes from their usage of more complex 3D shape representations. Qi et al. (2018) work with point clouds while Chen et al. (2019) sample 2D views from various viewpoints. Our approach, while simple in nature, provides competitive results compared with the current state-of-the-art in many-shot sketch-based 3D shape retrieval.

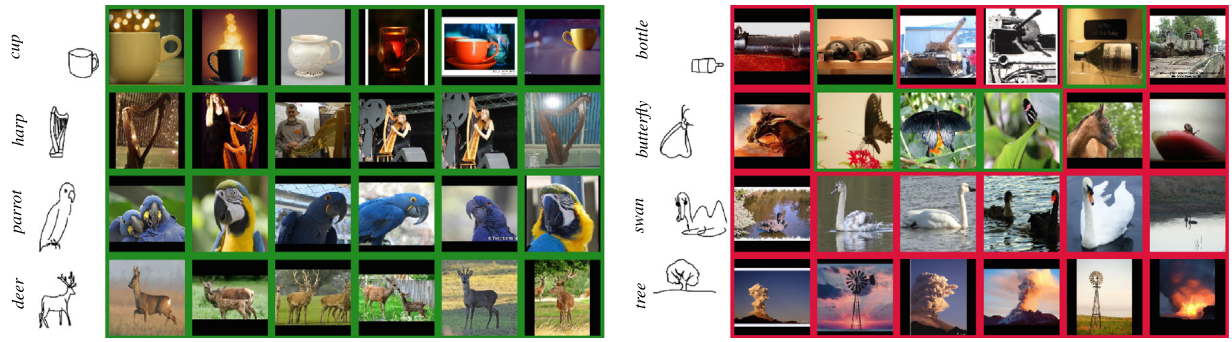
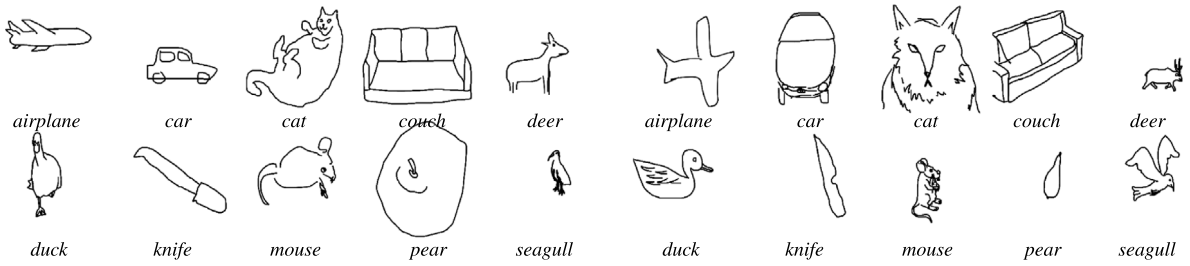


Fig. 9. Qualitative analysis of zero-shot sketch-based image retrieval. We show eight sketches of Sketchy Extended, with correct retrievals in *green*, incorrect in *red*. For typical sketches (*e.g.*, "cup"), the closest images are from the same category. For ambiguous sketches (*e.g.*, "tree") or non-canonical views (*e.g.*, "butterfly"), our approach struggles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Most effective set of sketches (86.07% accuracy). (b) Least effective set of sketches (43.82% accuracy).

Fig. 10. Qualitative analysis of few-shot sketch-based image classification on a subsampled Sketchy Extended. (a) 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.*, "deer" or "couch"). (b) The accuracy decreases when sketches are drawn badly (*e.g.*, "airplane"), or in non-canonical views (*e.g.*, "car" or "cat").

Table 5
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.

(a) SHREC13						
	NN	FT	ST	E	DCG	mAP
Siamese (Wang et al., 2015)	0.405	0.403	0.548	0.287	0.607	0.469
Shape2Vec (Tasse and Dodgson, 2016)	0.620	0.628	0.684	0.354	0.741	0.650
DCML (Dai et al., 2017)	0.650	0.634	0.719	0.348	0.766	0.674
LWBR (Xie et al., 2017)	0.712	0.725	0.785	0.369	0.814	0.752
DCA (Chen and Fang, 2018)	0.783	0.796	0.829	0.376	0.856	0.813
SEM (Qi et al., 2018)	0.823	0.828	0.860	0.403	0.884	0.843
DSSH (Chen et al., 2019)	**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**
(b) SHREC14						
	NN	FT	ST	E	DCG	mAP
Siamese (Wang et al., 2015)	0.239	0.212	0.316	0.140	0.496	0.228
Shape2Vec (Tasse and Dodgson, 2016)	0.714	0.697	0.748	0.360	0.811	0.720
DCML (Dai et al., 2017)	0.272	0.275	0.345	0.171	0.498	0.286
LWBR (Xie et al., 2017)	0.403	0.378	0.455	0.236	0.581	0.401
DCA (Chen and Fang, 2018)	0.770	0.789	0.823	0.398	0.859	0.803
SEM (Qi et al., 2018)	**0.804**	0.749	0.813	0.395	0.870	0.780
DSSH (Chen et al., 2019)	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**
(c) Part-SHREC14						
	NN	FT	ST	E	DCG	mAP
Siamese (Wang et al., 2015)	0.118	0.076	0.132	0.073	0.400	0.067
SEM (Qi et al., 2018)	**0.840**	0.634	0.745	0.526	0.848	0.676
DSSH (Chen et al., 2019)	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**

Qualitative analysis. To gain insight in our approach for retrieving 3D shapes from sketches, we provide qualitative examples in Fig. 11. Rotations of unaligned shapes can be handled. For example, 3D shapes of "laptop" or "piano" are retrieved despite the large differences in rotation angles. Yet, confusion remains with visually similar categories. This happens when the search needs to differentiate among fine-grained categories. For example, differences are subtle between "sedan cars" and "sports cars", or between "violin" and "cello". Although errors can appear with semantically similar categories, our method can retrieve highly variable 3D shapes from sketches.

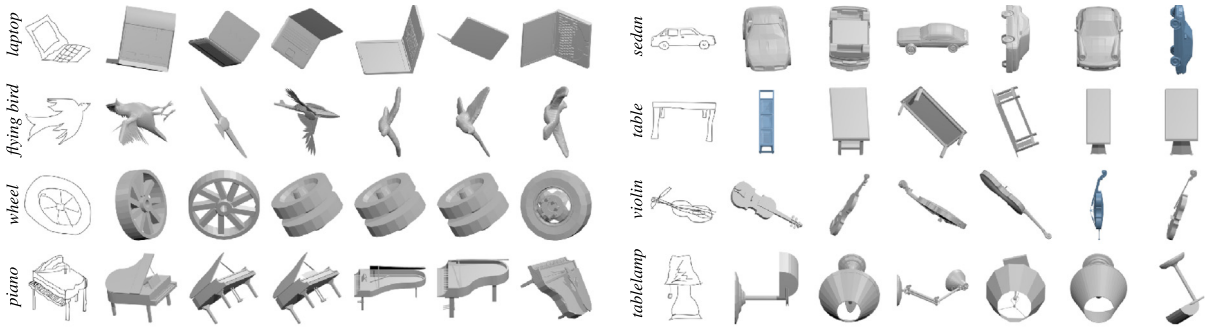


Fig. 11. Qualitative analysis of many-shot sketch-based 3D shape retrieval on Part-SHREC14. Incorrect results are shown in *blue*. Our approach handles the unaligned shapes by projecting all views to the same semantic prototype in the shared space. An open problem remains the confusion with categories that are close both in semantics and in appearance (*e.g.*, "violin" *vs.* "cello"). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

In this paper, we open visual search beyond two domains to scale to 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. This creates new challenges as all domains should map to the same embedding space, while new domains should be able to be incorporated efficiently. To achieve open cross-domain visual search, we propose a simple approach based on domain-specific prototype learners to align the semantics of multiple visual domains in a common space. Learning a mapping to a common space enables a visual search among any number of source or target domains. The addition of new domains consists in the training of a new prototype learner, without the need to retrain previous models. Empirical demonstrations on novel *open* cross-domain visual search tasks present how to search across multiple domains. State-of-the-art results on existing *closed* cross-domain visual search tasks show the effectiveness of our approach.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

William Thong is partially supported by a scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S., 2009. Diversifying search results. In: ACM WSDM.
Blanchard, G., Lee, G., Scott, C., 2011. Generalizing from several related classification tasks to a new unlabeled sample. In: NeurIPS.
Bui, T., Ribeiro, L., Ponti, M., Collomosse, J., 2017. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. CVIU 164.
Bui, T., Ribeiro, L., Ponti, M., Collomosse, J., 2018. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. Comput. Graph. 71.
Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T., 2019. Domain generalization by solving jigsaw puzzles. In: CVPR.
Chen, J., Fang, Y., 2018. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In: ECCV.
Chen, J., Qin, J., Liu, L., Zhu, F., Shen, F., Xie, J., Shao, L., 2019. Deep sketch-shape hashing with segmented 3d stochastic viewing. In: CVPR.
Chintala, S., Ranzato, M., Szlam, A., Tian, Y., Tygert, M., Zaremba, W., 2017. Scale-invariant learning and convolutional networks. Appl. Comput. Harmon. Anal. 42 (1), 154–166.
Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: CVPR.
Csurka, G., 2017. Domain Adaptation in Computer Vision Applications. Springer.
Dai, G., Xie, J., Zhu, F., Fang, Y., 2017. Deep correlated metric learning for sketch-based 3d shape retrieval. In: AAAI.
Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition. In: CVPR.
Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.-Z., 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In: CVPR.
Dou, Q., Castro, D.C., Kamnitsas, K., Glocker, B., 2019. Domain generalization via model-agnostic learning of semantic features. In: NeurIPS.
Dutta, A., Akata, Z., 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: CVPR.
Dutta, T., Biswas, S., 2019. Style-guided zero-shot sketch-based image retrieval. In: BMVC.
Eitz, M., Hays, J., Alexa, M., 2012. How do humans sketch objects?. ACM TOG 31 (4).
Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M., 2010. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. TVCG 17 (11).
Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. Devise: A deep visual-semantic embedding model. In: NeurIPS.
Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. JMLR 17 (1).
Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F., 2012. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE TPAMI 35 (12).
Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y., 2018. Image-to-image translation for cross-domain disentanglement. In: NeurIPS.
Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: CVPR.
Hinton, G., Vinyals, O., Dean, J., 2014. Distilling the knowledge in a neural network. In: NeurIPS-W.
Hu, R., Collomosse, J., 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. CVIU 117 (7).
Hu, C., Li, D., Song, Y.-Z., Xiang, T., Hospedales, T.M., 2018a. Sketch-a-classifier: sketch-based photo classifier generation. In: CVPR.
Hu, J., Shen, L., Sun, G., 2018b. Squeeze-and-excitation networks. In: CVPR.
Hu, R., Wang, T., Collomosse, J., 2011. A bag-of-regions approach to sketch-based image retrieval. In: ICIP.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML.
Jacobs, C.E., Finkelstein, A., Salesin, D.H., 1995. Fast multiresolution image querying. In: SIGGRAPH.
Johnson, J., Douze, M., Jégou, H., 2017. Billion-scale similarity search with gpus. arXiv:1702.08734.
Kato, T., 1992. Database architecture for content-based image retrieval. In: Image Storage and Retrieval Systems, Vol. 1662.
Li, B., Lu, Y., Godil, A., Schreck, T., Aono, M., Johan, H., Saavedra, J.M., Tashiro, S., 2013. Shrec'13 track: large scale sketch-based 3d shape retrieval. In: Eurographics Workshop on 3D Object Retrieval.
Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M.J., Johan, H., Matsuda, T., et al., 2014a. A comparison of methods for sketch-based 3d shape retrieval. CVIU 119.
Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al., 2014b. Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In: Eurographics Workshop on 3D Object Retrieval.
Li, D., Yang, Y., Song, Y.-Z., Hospedales, T.M., 2017. Deeper, broader and artier domain generalization. In: ICCV.
Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L., 2017a. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: CVPR.
Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017b. Sphereface: Deep hypersphere embedding for face recognition. In: CVPR.

Liu, Q., Xie, L., Wang, H., Yuille, A., 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In: ICCV.
Loshchilov, I., Hutter, F., 2017. Sgdr: Stochastic gradient descent with warm restarts. In: ICLR.
Lu, P., Huang, G., Fu, Y., Guo, G., Lin, H., 2018. Learning large euclidean margin for sketch-based image retrieval. arXiv:1812.04275.
Mensink, T., Verbeek, J., Perronnin, F., Csurka, G., 2013. Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE TPAMI 35 (11).
Mettes, P., van der Pol, E., Snoek, C.G., 2019. Hyperspherical prototype networks. In: NeurIPS.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: ICLR.
Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S., 2017. No fuss distance metric learning using proxies. In: ICCV.
Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M., 2009. Zero-shot learning with semantic output codes. In: NeurIPS.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS.
Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B., 2019. Moment matching for multi-source domain adaptation. In: ICCV.
Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K., 2017. Visda: The visual domain adaptation challenge. arXiv:1710.06924.
Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation. In: EMNLP.
Phong, B.T., 1975. Illumination for computer generated pictures. Commun. ACM 18 (6), 311–317.
Qi, A., Song, Y.-Z., Xiang, T., 2018. Semantic embedding for sketch-based 3d shape retrieval. In: BMVC.
Qi, Y., Song, Y.-Z., Zhang, H., Liu, J., 2016. Sketch-based image retrieval via siamese convolutional neural network. In: ICIP.
Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: LREC Workshop on New Challenges for NLP Frameworks. ELRA.
Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S., 2018. Meta-learning for semi-supervised few-shot classification. In: ICLR.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. IJCV 115 (3).
Saavedra, J.M., 2014. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In: ICIP.
Sangkloy, P., Burnell, N., Ham, C., Hays, J., 2016. The sketchy database: Learning to retrieve badly drawn bunnies. ACM TOG.
Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: CVPR.
Shen, Y., Liu, L., Shen, F., Shao, L., 2018. Zero-shot sketch-image hashing. In: CVPR.
Shilane, P., Min, P., Kazhdan, M., Funkhouser, T., 2004. The Princeton shape benchmark. In: Shape Modeling International.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: ICLR.
Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: NeurIPS.
Sohn, K., 2016. Improved deep metric learning with multi-class n-pair loss objective. In: NeurIPS.
Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.G., 2015. Multi-view convolutional neural networks for 3d shape recognition. In: ICCV.
Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: ICML.
Tasse, F.P., Dodgson, N., 2016. Shape2vec: Semantic-based descriptors for 3d shapes, sketches and images. ACM TOG.
Wang, F., Kang, L., Li, Y., 2015. Sketch-based 3d shape retrieval using convolutional neural networks. In: CVPR.
Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. Cosface: Large margin cosine loss for deep face recognition. In: CVPR.
Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. 10.
Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. In: ECCV.
Wilber, M.J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., Belongie, S., 2017. Bam! the behance artistic media dataset for recognition beyond photography. In: ICCV.
Xian, Y., Lampert, C.H., Schiele, B., Akata, Z., 2018. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE TPAMI.
Xie, J., Dai, G., Zhu, F., Fang, Y., 2017. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In: CVPR.
Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L., 2018. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: CVPR.
Yelamarthi, S.K., Krishna Reddy, S., Mishra, A., Mittal, A., 2018. A zero-shot framework for sketch based image retrieval. In: ECCV.
Zhai, A., Wu, H.-Y., 2019. Making classification competitive for deep metric learning. In: BMVC.
Zhang, H., Liu, S., Zhang, C., Ren, W., Wang, R., Cao, X., 2016. Sketchnet: Sketch classification with web images. In: CVPR.
Zhuo, J., Wang, S., Cui, S., Huang, Q., 2019. Unsupervised open domain recognition by semantic discrepancy minimization. In: CVPR.