Open Cross-Domain Retrieval without Examples

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

Abstract

This paper strives for cross-domain zero-shot retrieval, where unseen categories in a target domain are retrieved with queries from a source domain. Such a cross-domain retrieval has so far focused on a closed setting between two pre-defined domains. We open the retrieval to any number of source and target domains. To achieve this, prototype learners map every domain to a common semantic space. We learn domain-specific models to regress inputs to their corresponding categorical prototype in the embedding space. By doing so, we create a common semantic space where the open cross-domain retrieval occurs. We evaluate the proposed methodology in an open setting where we retrieve categories from any source to any target domains. In the traditional closed setting from one source to one target domain, we achieve state-of-the-art results on the well-established sketch-based image retrieval task.

1. Introduction

This paper strives for the cross-domain zero-shot retrieval problem, where unseen categories in a target domain are retrieved with queries from a source domain. An active line of works has notably focused on retrieving natural images of unseen categories from sketch queries [1–3, 8, 10, 20]. In their context, individual source and target domains are known a priori. We propose to open the retrieval problem to multiple domains. For example, we may want to on-the-fly use clipart of the "Space Needle" tower in Seattle to retrieve 3D shapes of it, or draw a sketch of the tower to retrieve examples from cliparts. Enabling cross-domain retrieval for open 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.

Inspired by recent works on prototype-based embedding spaces [12, 17, 19], we introduce a prototype learner to map every domain to a common and shared semantic space. 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 space. As a direct advantage over existing works, the proposed formulation removes the need for domain adaptation [1–3, 10, 20] or knowledge preservation [8] losses to create a domain-agnostic embedding space. Indeed in our approach, domain-specific models all map to the same semantic space where cross-domain retrieval occurs. By relying on a prototype learner, joint training of multiple models or intricate sampling among domains are no longer required [12, 19]. And by training separate models, adding a new domain amounts to training a new domain-specific model only, without the need to retrain the existing models.

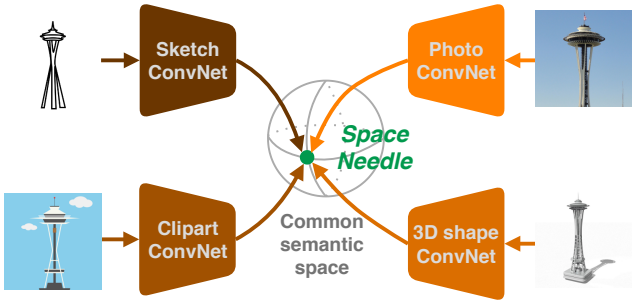

Figure 1. **Open cross-domain zero-shot retrieval.** We retrieve images of unseen categories from any source domain to any target domain, *e.g.* a retrieval of the "Space Needle" tower in Seattle from cliparts to 3D shapes, or from sketches to cliparts.

Our key contribution is the introduction of open cross-domain zero-shot retrieval, where unseen categories are retrieved between any domain pair. We introduce a simple approach based on the mapping to a common space with categorical prototypes (Section 2). We evaluate the methodology for retrieving unseen categories in an open setting from *any* source domain to *any* target domain (Section 3.1). We also show the effectiveness of our approach in a closed setting from *one* source domain to *one* target domain on the well-established zero-shot sketch-based image retrieval task, where we achieve state-of-the-art results (Section 3.2).

2. Method

Consider the example in Figure 1, where the "Space Needle" tower in Seattle can be visualized and stored in four different manners. The cross-domain task is to retrieve visuals of the "Space Needle", from any source domain to any target domain. This can take the form of going from cliparts to 3D shape, or from sketches to cliparts. To achieve this, the idea is to map "Space Needle" visuals from any domain to the same common semantic space. In other words, we want the representation of these visuals from any domain to be close to the representation of the category "Space Needle" in the embedding space. We propose to learn of a separate mapping function for each domain with a prototype-based loss function.

Problem formulation Let $\mathcal{T}_d = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N_d}$ be a visual collection from domain $d \in \mathcal{D}$ with N_d samples. $\mathbf{x}^{(n)}$ denotes an input example of category $y^{(n)} \in \mathcal{Y}$. As we consider retrieval in a zero-shot setting, the category vocabulary $\mathcal{Y}$ is split into $\mathcal{Y}_{tr}$ and $\mathcal{Y}_{gal}$ for training and evaluation respectively, with $\mathcal{Y}_{tr} \cup \mathcal{Y}_{gal} = \emptyset$. Hence, categories to be retrieved in the gallery set during evaluation have not been seen during training.

Mapping function 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. The mapping function is formulated as a convolutional network (ConvNet) with ℓ_2-normalization on the D-dimensional network outputs, which creates a hyperspherical embedding space $\mathbb{S}^{D-1}$.

Categorical prototypes Given a visual collection $\mathcal{T}_d$, we propose a probabilistic model to map an example $\mathbf{x}$ from domain d to the shared semantic space:

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

where $\phi(y)$ is the corresponding categorical prototype, $c(\cdot, \cdot)$ the cosine distance and $s \in \mathbb{R}_{>0}$ the scaling factor. The scaling factor, inversely equivalent to the temperature [5], controls how samples are spread around categorical prototypes. In this work, we define the shared semantic space with word embeddings to represent categories. Intuitively, the probabilistic model pulls $f_d(\mathbf{x})$ to its corresponding semantic prototype $\phi(y)$ at each step, and pushes it away from the prototypes of other categories $\phi(y')$.

Finally, learning every mapping function f_d is done by minimizing the cross-entropy over the training set:

$$\mathcal{L}_d = -\frac{1}{N_d} \sum_{n=1}^{N_d} \log p(y^{(n)}|\mathbf{x}^{(n)}, d). \quad (2)$$

Retrieval Once trained, the categorical prototypes are discarded. In the retrieval phase, we compute the cosine similarity between the source query and target samples in the shared semantic space. This provides a rank for all target samples with respect to the query.

Query refinement As there can exist an inherent variability in the hyperspherical semantic space that can cause noise in the similarity measures, 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:

$$\hat{p}(p_0, p_1|\lambda) = \frac{\sin\big((1-\lambda)\Omega\big)}{\sin\Omega} p_0 + \frac{\sin\big(\lambda\Omega\big)}{\sin\Omega} p_1, \quad (3)$$

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 neighbour of p_0 in the target set. This mixture doesn't require any label and relies on the fact that the recall at one is usually very high.

3. Results

We evaluate our method quantitatively on a new task which retrieves categories from any source to any target domain (Section 3.1), and on a well-established task in the literature which retrieves categories from one source to one target domain (Section 3.2).

Implementation details Throughout this paper, we use SE-ResNet50 [6] pre-trained on ImageNet [14] as a backbone network, and word2vec trained on a Google News corpus [11] as the common semantic space. We remove the final classifier layer of SE-ResNet50, and replace it with a fully-connected layer of size $D = 300$ with random weights. We rely on Nesterov momentum [18] for optimization with a coefficient to 0.9, a learning rate of $1\mathrm{e}{-4}$ with cosine annealing without warm restarts [9] and a batch size of 128. We use a scaling factor s of 20 and set $\lambda = 0.7$.

3.1. From any source to any target domain

Setup We evaluate on the recently introduced *DomainNet* [13], which contains 596,006 images from 345 classes. Images are gathered from six visual domains: *clipart*, *infograph*, *painting*, *pencil*, *photo* and *sketch*. We split samples into 300 training and 45 testing classes with at least 40 samples per class and report the mean average precision over the whole gallery set (mAP@all).

Results We demonstrate how retrieving from any source to any target domain in an open setting is enabled by our approach. Figure 2 shows the result of 36 cross-domain zero-

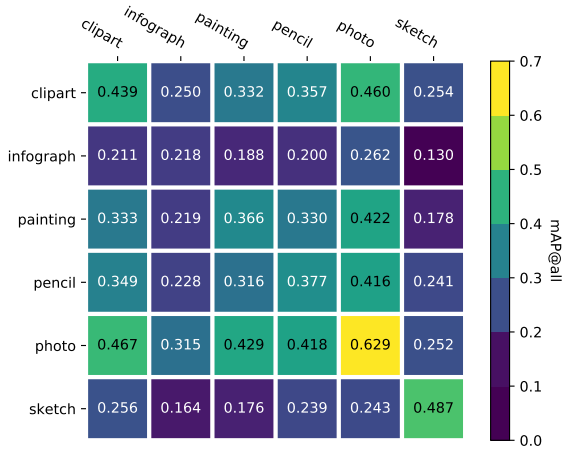

Figure 2. **From any source to any target domain.** Our approach can perform 36 cross-domain retrievals in a zero-shot evaluation on 45 unseen classes by only training 6 models. Rows refer to the source domain while columns to the target domain. Searching from and within the photograph domain is the most effective while infographs are the most challenging. We report the mAP@all.

Figure 3. **Retrieval examples** on DomainNet. Given a sketch query from the "cat" category, the top-6 images from the five other domains are retrieved. Correct results are in *green*, errors in *red*. The retrieval within the photograph domain is the most effective while errors appear in other domain mainly because of semantics relatedness with other categories (*e.g.* "tiger" or "racoon").

shot retrieval evaluations, corresponding to all six cross-domain pairs. In our formulation, only six models are required to be trained, one for every domain. For comparison, a domain adaptation approach –the standard in current cross-domain retrieval methods– requires a pair-wise training of all available domain combinations. While approaches based on pair-wise training scale with a quadratic complexity to the number of domains, we scale linearly.

Quantitatively, the photograph domain provides the most effective retrieval 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, infograph and sketch domains induce a much more difficult retrieval. This comes from the fact that both are very diverse in terms of scale and appearance.

Figure 3 illustrates the retrieval task where a single sketch query targets galleries from different domains. This qualitative example confirms that retrieving within the photo domain is the most effective as cats are well defined in images. On the contrary, infograph domain depict a high variability as cats exhibit various sizes or forms and images are over-cluttered with noisy visuals.

3.2. From one source to one target domain

Setup We evaluate on the zero-shot sketch-based image retrieval, which focuses on retrieving natural images from a sketch query. We compare on two datasets. *TU-Berlin Extended* [4,21] contains 20,000 sketches and 204,070 images from 250 classes. Following Shen *et al*. [16], we select 220 classes for training and 30 classes for evaluation. *Sketchy Extended* [7, 15] contains 75,481 sketches and 73,002 images from 125 classes. SImilarly, following Shen *et al*. [16], we select 100 classes for training and 25 classes for evaluation. For fair comparison with Liu *et al*. [8], we select the same unseen classes for both datasets and pick the same backbone network SE-Resnet50. Following recent works [2, 8, 16], we report the mAP@all and the precision at 100 (prec@100) scores.

Table 1. **From sketches to images** on Sketchy Extended and TU-Berlin Extended. Aligning solely the semantics improves cross-domain image retrieval.

	TU-Berlin Extended		**Sketchy Extended**	
	mAP@all	prec@100	mAP@all	prec@100
EMS [10]	0.259	0.369	n/a	n/a
CAAE [20]	n/a	n/a	0.196	0.284
ADS [1]	0.110	n/a	0.369	n/a
SEM-PCYC [2]	0.297	0.426	0.349	0.463
SG [3]	0.254	0.355	0.376	0.484
SAKE [8]	0.475	**0.599**	0.547	0.692
This paper	**0.517**	0.557	**0.649**	**0.708**

Results Table 1 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. 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*. [8] obtains a higher prec@100 score. 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.

Figure 4 provides several qualitative example sketches with their top-6 retrieved images. Our approach works well for sketches of categories with a canonical view and a typi-

Figure 4. **Qualitative analysis** of zero-shot sketch-based image retrieval. We show six sketches of Sketchy Extended, with correct retrievals in *green*, incorrect in *red*. For ambiguous sketches (*e.g.*, tree) or non-canonical views (*e.g.*, butterfly), our approach struggles. For typical sketches (*e.g.*, cup), the closest images are from the same category.

cal appearance. Results degrade when sketches are in non-canonical views or ambiguous.

4. Conclusion

In this paper, we present open cross-domain zero-shot retrieval from any source domain to any target domain. In contrast to previous work, retrieving categories is no longer restricted to a search between two domains. We introduce a prototype-based learner for every domain. During training, every domain-specific model regresses inputs to their corresponding categorical prototype. This simple approach allows to scale the retrieval to any number of domains, and to add new domains without hassle. Evaluation on the closed setting with a well-established task shows the effectiveness of the method.

Acknowledgments William Thong is partially supported by an NSERC scholarship.

References

[1] Sounak 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, 3

[2] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 1, 3

[3] Titir Dutta and Soma Biswas. Style-guided zero-shot sketch-based image retrieval. BMVC, 2019. 1, 3

[4] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 31(4), 2012. 3

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS-W*, 2014. 2

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2

[7] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 3

[8] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, 2019. 1, 3

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 2

[10] Peng Lu, Gao Huang, Yanwei Fu, Guodong Guo, and Hangyu Lin. Learning large euclidean margin for sketch-based image retrieval. *arXiv:1812.04275*, 2018. 1, 3

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 2

[12] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *ICCV*, 2017. 1

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

[14] 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. 2

[15] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 3

[16] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 3

[17] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 1

[18] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013. 2

[19] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 1

[20] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 1, 3

[21] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. 3