Supplementary Material: Towards Unsupervised Sketch-based Image Retrieval

Conghui Hu[1]
conghui@nus.edu.sg
Yongxin Yang[2]
yongxin.yang@ed.ac.uk
Yunpeng Li[3]
yunpeng.li@surrey.ac.uk
Timothy M. Hospedales[2]
t.hospedales@ed.ac.uk
Yi-Zhe Song[3]
y.song@surrey.ac.uk

[1] National University of Singapore
Singapore
[2] University of Edinburgh
United Kingdom
[3] University of Surrey
United Kingdom

1 Algorithm

Algorithm 1: Unsupervised SBIR training

Input:
Sketches $\mathcal{I}^s$; Photos $\mathcal{I}^p$;
Output:
Feature extractor f_θ

repeat
  Randomly select a mini-batch $\{I_i^s, I_i^p\}_{i=1}^A$;
  Update feature queue $\mathbf{Q}$, sketch memory bank $\mathbf{M}^s$ and photo memory bank $\mathbf{M}^p$
  Fix feature extractor f_θ and prototypes $\mathbf{U}$, and solve for $\hat{\Gamma}^s$ and $\hat{\Gamma}^p$ as in Equation 1;
  Fix $\hat{\Gamma}^s$ and $\hat{\Gamma}^p$, and update feature extractor f_θ and prototypes $\mathbf{U}$ according to Equation 6;
until Convergence or max training iterations

2 Models for comparison

Unsupervised SBIR We compare our method with the following unsupervised representation learning methods: **RotNet [4]** A self-supervised method that uses rotation prediction as the pretext task. Here, we perform 4-class (0°, 90°, 180°, 270°) classification. **ID [10]** Instance-discrimination for unsupervised representation learning, ignoring image domain. **CDS [6]** A self-supervised cross-domain method that performs intra-domain instance discrimination and cross-domain matching for representation learning. **GAN [5]** Adversarial

learning is enabled by introducing a 3-layer MLP network as the discriminator to distinguish extracted features from both domains. While the feature extractor learns with the objective of removing domain-dependent feature and fooling the discriminator. **SwAV [1]** An unsupervised feature representation learning algorithm where the cluster assignments for different variants of the same image are enforced to be consistent. During the training procedure, sketches and photos are mixed together ignoring the domain differences. **DSM [8]** A feature extractor is trained with contrastive loss leveraging matching and non-matching training edgemap pairs. DSM results are generated by the pre-trained model from the original paper. **SwAV [1] + CycleGAN [12]** CycleGAN, which achieves unpaired image translation, is first trained with unlabeled data to convert sketch to color photos. Both real photo data and images generated from sketch are then used in SwAV. **SwAV [1] + GAN [5]** Different from GAN only method, feature extractor now aims at generating both semantic-aware and domain-agnostic feature by combining the swapping label prediction and adversarial learning together.

Zero-shot SBIR Here, we compare with recent advances in zero-shot SBIR. Note that the other competitors leverage sketch-photo pairing or class label annotation to train the model while ours does not. **ZSIH [9]** Sketch and photo are encoded into binary codes for retrieval. Sketch-photo heterogeneity is remedied by Kronecker fusion layer, graph convolution and word embeddings. **CVAE [11]** Variational autoencoder is employed to model the probability distribution over images conditioned on its paired sketch feature and generate a latent prior vector. The sketch feature along with latent prior is then projected onto photo feature space. **SAN [7]** A multi-staged generative model is designed to transform sketch feature and ensures features from different domains are encoded into a common subspace. **SEM-PCYC [3]** Cycle-consistency is used to map data from both domains to a shared semantic space while preserving the ability to translate back to the original modality. **Doodle [2]** Domain loss and triplet ranking loss are used to learn a common embedding space where distance between instance pairs in class-wise alignment are smaller than unaligned pairs.

3 Further Analysis

Influence of memory bank size To investigate the impact of our new memory bank on optimal transport-aided domain alignment, we adjust the memory bank size and show the results in Table 1. We can see that: (i) the use of memory banks indeed improves the efficacy of OT, even with a small size as 480; (ii) retrieval accuracy improves gradually with memory bank size before saturating at around size 4000.

Table 1: Memory bank size hyperparameter sensitivity. Unsupervised SBIR results on Sketchy-Extended dataset.

Mem. bank size	Prec@200(%)	mAP@200 (%)	mAP (%)
no	25.60	28.62	20.98
480	27.15	30.32	22.80
960	30.36	33.54	25.59
1920	33.02	35.93	27.64
3840	**33.64**	**36.31**	**28.17**

Influence of prototype number We evaluate whether the number of prototypes has an effect on the unsupervised cross-domain retrieval. The results in Figure 1 show that (i)

by matching the prototype number with training category size (125 categories in Sketchy-Extended dataset), the model performs the best, so taking the class number as a known condition assists parameter optimization; (ii) slightly increasing the prototype number to 150 still leads to comparable retrieval results.

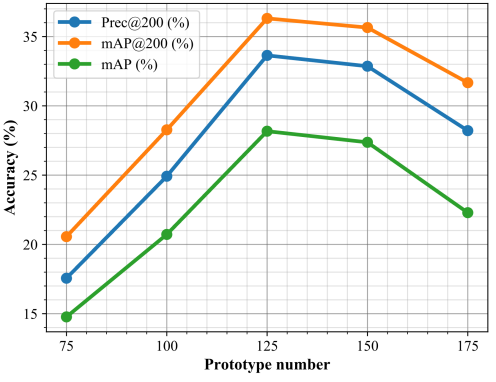

Figure 1: Prototype number K hyperparameter sensitivity. Unsupervised SBIR results on Sketchy-Extended dataset.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2020. URL http://arxiv.org/abs/2006.09882. arXiv:2006.09882.

[2] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2179–2188, 2019.

[3] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5089–5098, 2019.

[4] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2020. URL http://arxiv.org/abs/1803.07728. arXiv:1803.07728.

[5] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL http://arxiv.org/abs/1406.2661. arXiv:1406.2661.

[6] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels, 2020. URL http://arxiv.org/abs/2003.08264. arXiv:2003.08264.

[7] Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, Anurag Mittal, and Hema Murthy. Stacked adversarial network for zero-shot sketch based image retrieval. In *Proc. IEEE/CVF Winter Conf. on Appl. of Comput. Vis. (WACV)*, pages 2540–2549, 2020.

[8] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 751–767, 2018.

[9] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3598–3607, 2018.

[10] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3733–3742, 2018.

[11] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proc. Euro. Conf. on Comput. Vis. (ECCV)*, pages 300–317, 2018.

[12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 2223–2232, 2017.