Supplementary: Data-Free Sketch-Based Image Retrieval

Abhra Chaudhuri
University of Exeter, UK
ac1151@exeter.ac.uk

Ayan Kumar Bhunia, Yi-Zhe Song, Anjan Dutta
Institute for People-Centred AI, University of Surrey, UK
{a.bhunia, y.song, anjan.dutta}@surrey.ac.uk

1. Experimental Details

Datasets: The Sketchy dataset contains 75,471 sketches and 12,500 photos, with 60,502 additional photos from ImageNet [1] in its extension [5], evenly distributed over 125 classes, with both class and instance level correspondences. The TU-Berlin dataset contains 20,000 sketches evenly distributed over 250 categories, with a total of 204,489 class-level natural image correspondences in its extension [5]. The QuickDraw-Extended dataset contains 330,000 sketches and 204,000 photos from 110 categories. Following existing literature [2, 5], we use 10 and 50 randomly selected sketches per category from TU-Berlin and Sketchy respectively for testing the trained encoders in the category level SBIR setting, with the remaining sketches and photos used for training the teacher classifiers. We follow the same procedure as that of TU-Berlin for experimenting with the QuickDraw-Extended dataset.

Pre-Trained Classifiers: We trained ResNet50 [3] models on the train splits of Sketchy and TU-Berlin to obtain the photo and sketch classifier networks that would act as teachers. We used Adam as the optimizer with a learning rate of 0.01 under an exponential decay rate of 0.98, and weight decay of 10^{-5}. The photo and sketch classifiers were trained up to accuracies of 96.34% and 93.81% for Sketchy, and 92.70% and 81.35% for TU-Berlin respectively.

Implementation Details: Our estimator networks follow the architecture of the StyleGAN 2 [4] generator, trained using the Adam optimizer with a learning rate of 0.02. Our encoders have a ResNet50 [3] backbone, also trained using the Adam optimizer with a learning rate of 2×10^{-3}, decayed using a cosine annealing schedule. We initially train the estimators for 100 epochs in a warm-up phase for the estimated samples to stabilize and approach closer to the ones belonging to the true distribution. Thereafter, we train both the estimator and the generator pairs in an alternating manner (each frozen while the other is updated) for 500 epochs. In each epoch, we generate 10,000 positive pairs of photo-sketch reconstructions.

Platform Details: We implement our data-free SBIR pipeline on an Ubuntu 20.04 workstation with a single NVIDIA RTX 3090 GPU, an 8-core Intel Xeon processor and 32 GBs of RAM, using the PyTorch [7] deep learning framework. By the virtue of using a fixed-size, gradient-free queue for storing negative instances for contrastive learning of the encoders, our method bypasses dependencies on the batch-size, thus allowing us to perform the training end-to-end on a single GPU.

1.1. Additional Details on Baselines

Sampling from a Gaussian Prior: The input photos and sketches constitute samples drawn from an $N \times N$ Gaussian distribution, where N is the expected input spatial dimension for the downstream encoder. We assign labels to such samples in a manner that ensures equal number of samples across all classes. We then use these samples for training the encoders.

Averaging Weights: We speculate that averaging and using a single network for encoding photos and sketches helps bridge the modality gap. As has been demonstrated time and again in relevant literature, modality-specific, semantically irrelevant features is the single biggest source of error in SBIR. With averaging weights, we are able to use a single network for encoding photos and sketches, while incorporating the knowledge about modality specific variances learned by the individual networks.

Meta-Data Based Reconstruction: Following [6], we retain the means and the covariances of activations from all layers of the classifier. We then use them as metadata for input reconstruction, by generating samples that induce similar activation statistics across all layers of the classifier.

2. Qualitative Ablations

2.1. Class Alignment

The contribution of the Class-Alignment loss ($\mathcal{L}_{\text{align}}$) in reconstructing semantically matching photo-sketch pairs given a common noise vector ξ is visualized in Figure 1. It can be seen that without $\mathcal{L}_{\text{align}}$, the estimators often produce samples that belong to largely different classes. This sends wrong signals to the downstream encoders in terms of learning a semantically meaningful metric-space. With the class-alignment loss, the estimators can be guaranteed

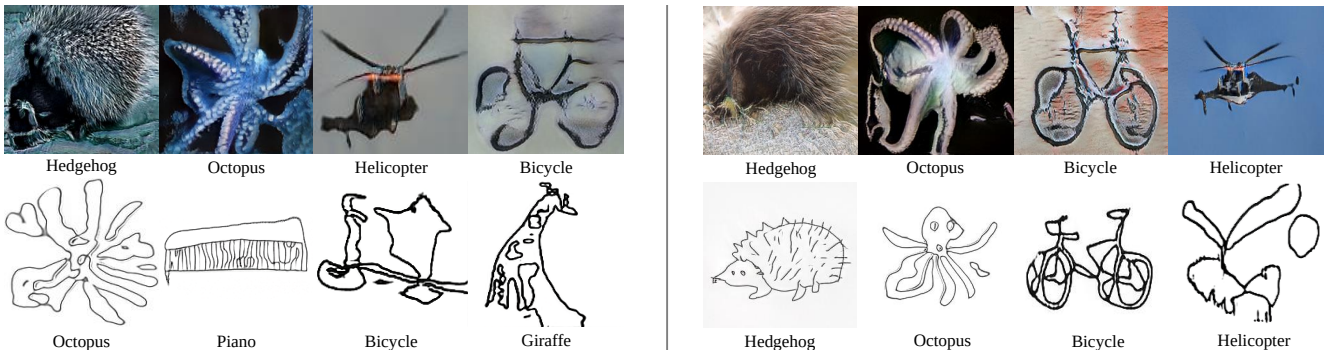
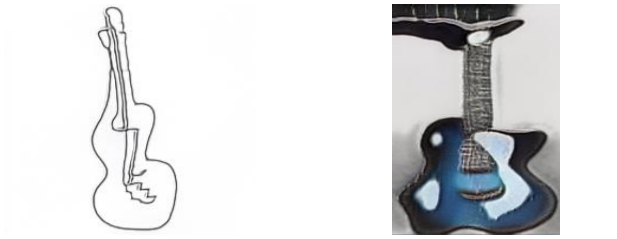


Figure 1. Photo and sketch reconstructions without (left) and with (right) the Class-Alignment loss ($\mathcal{L}_{\text{align}}$). Each column corresponds to a single reconstruction step using a common input noise vector ξ fed in to the photo and sketch estimators respectively.



Guitar: 37%, Violin: 35%

Guitar: 43%, Violin: 4%

Figure 2. Examples of reconstructions for which the output distributions of the sketch and photo classifiers differ significantly.

to receive photo-sketch pairs that have the same class information, and hence qualify as correct positive pairs for optimizing the encoders.

Also, as discussed in Section 3.1 of the main text, pairing solely on the basis of discrete labels (obtained via an argmax on the teachers' output) may not faithfully represent the semantic content in a reconstruction. Figure 2 shows such an example. Even though their hard-labels are the same (Guitar), the distribution of class information is vastly different in the two images. Thus, for such pairs, it is important for their predicted distributions to be properly aligned (via an objective like $\mathcal{L}_{\text{align}}$) before they can actually be considered as positive pairs for downstream metric learning.

2.2. Modality Guidance

Figure 4 qualitatively demonstrates the contribution of our Modality Guidance Network (d_ϕ), and its objective function $\mathcal{L}_{\text{modal}}$. It can be seen that Unguided estimators cannot make a clear distinction among the modalitites – photo reconstructions contain object outlines like sketches, and sketch reconstructions contain colors. This happens because in presence of the class-alignment loss ($\mathcal{L}_{\text{align}}$), the estimators exchange information across modalities. Under such a circumstance, the estimators can minimize both semantic distance, as well as modality distance in order to minimize $\mathcal{L}_{\text{align}}$. The task of our Modality Guidance Network is to ensure that the estimators only minimize $\mathcal{L}_{\text{align}}$ by minimizing semantic distance, and not through the exchange of modality-specific information. With this, the Modality Guided estimators are bounded in their output space, producing clean and realistic reconstructions.

2.3. Metric-Agnostic Adversarial Estimation

Figure 5 shows sample reconstructions obtained by optimizing our Metric-Agnostic Adversarial Estimation $\mathcal{L}_{\text{adv}}$ loss. While the teachers predict them to be instances of 'Piano' with very high confidence while assigning the class of 'Bathtub' a low probability, the predictions from the students is just of the opposite nature. This makes such samples the hardest for the student to encode, as their predictions are highly divergent from those of the teachers. Optimizing on such samples thus makes the student more robust to challenging real-world test cases.

3. Reconstruction Quality and Performance

Figure 3 shows the relationship between the quality of the reconstructed samples and the mean average precision (mAP) of the encoders on the Sketchy datasets. Training the estimators for longer helps in the reconstruction of more realistic samples. However, beyond a certain point, realism does not seem to significantly affect the retrieval performance. As the estimators start capturing the fundamental shape and texture of the object, there is a significant improvement in mAP from 0.45 (epoch 300) to 0.68 (epoch 350). With a few more epochs of training, the accuracy steadily increases to 0.77 (epoch 400). Beyond this point, the increase is much slower, although the quality of the reconstructions keep getting better.

4. Training with Partial Class Overlap

For the classes that are unknown to the photo classifier, we randomly initialize *trainable* proxy-vectors to act

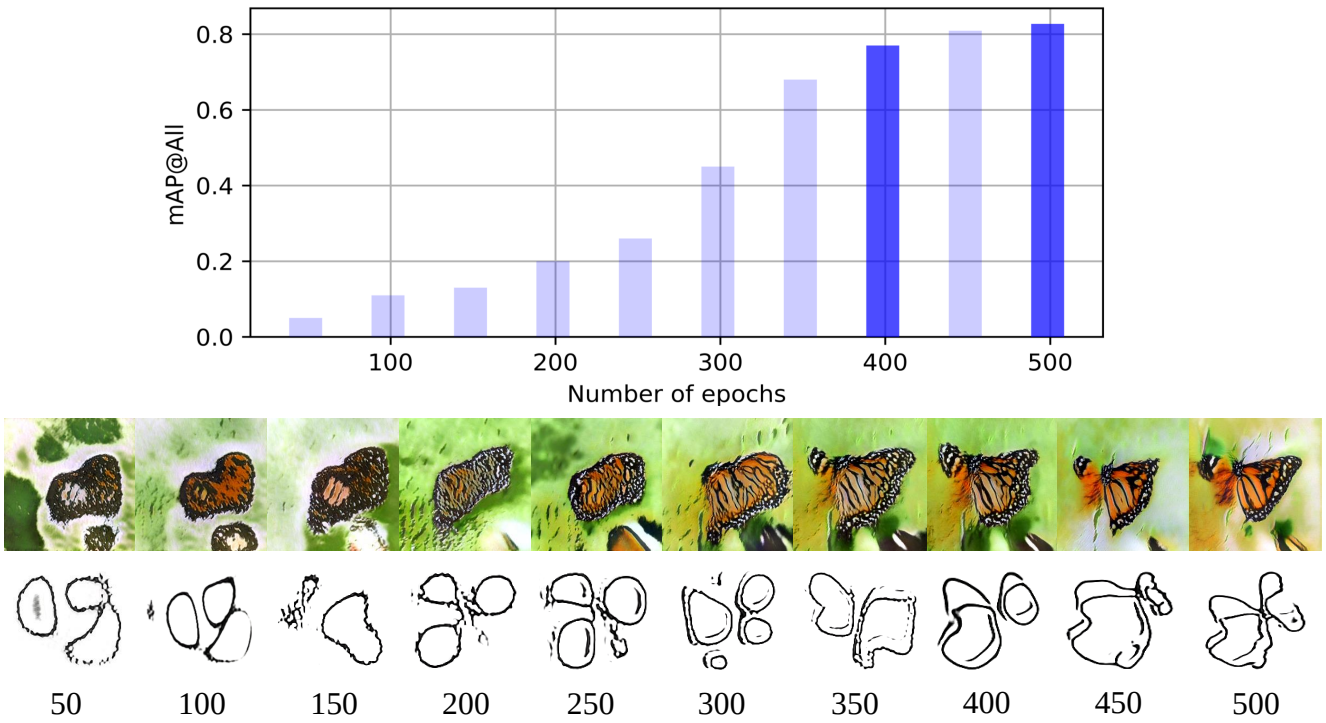


Figure 3. Variation in mAP@all, as well as reconstruction quality across epochs.

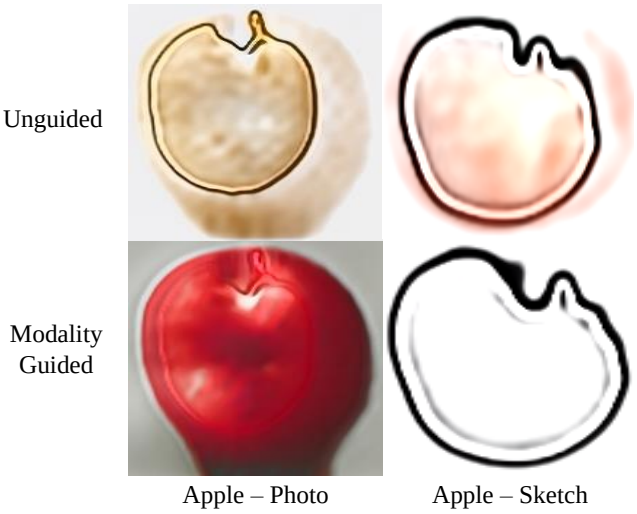


Figure 4. Reconstructed photos and sketches of an Apple in the presence and absence of the Modality Guidance loss ($\mathcal{L}_{\text{MG}}$).

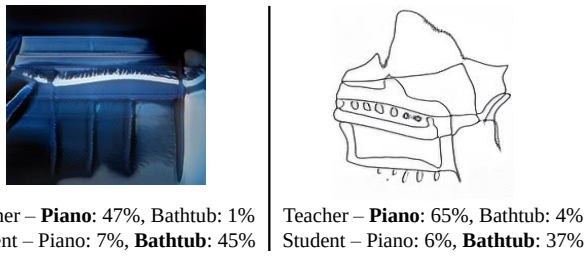


Figure 5. Sample reconstructions obtained by using our Metric-Agnostic Adversarial Estimation ($\mathcal{L}_{\text{adv}}$) criterion, with respective class-scores assigned by the teacher and the student.

as representatives for those classes, and concatenate them to the final layer neurons of the classifier. We do the same for the sketch classifier, and reorder the proxy vectors in both modalities to have consistency in class indices across modalities. For each reconstruction, if the sketch belongs to a class that is unknown to the photo classifier, we consider the prediction from the sketch classifier to be the ground-truth of the corresponding photo. We update the trainable proxies in the final layer of the photo classifier based on this information. We do a symmetric operation for the sketch classifier as well. Note that such alterations to the teachers are only possible because the reconstructed photos and sketches have been semantically aligned by our $\mathcal{L}_{\text{align}}$ objective. The rest of the process is performed as usual.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009. 1

[2] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020. 1

[3] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 1

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1

[5] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 1

[6] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv*, 2017. 1

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPSW*, 2017. 1