Supplementary material for Sketching *without* Worrying: Noise-Tolerant Sketch-Based Image Retrieval

Ayan Kumar Bhunia[1] Subhadeep Koley[1,2] Abdullah Faiz Ur Rahman Khilji*
Aneeshan Sain[1,2] Pinaki nath Chowdhury [1,2] Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{a.bhunia, s.koley, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

1 Comparative Study with different RL methods

We compare with different RL methods [5, 1], starting from Vanilla Policy Gradient, Deep Q-Learning, TRPO, to variants of PPO. For our use-case we get best results (Table 1) with PPO actor-critic version with clipped surrogate objective, where the critic network leads to one important byproduct of modelling retrieval ability of partial sketches.

Table 1: Performance analysis using different RL approaches.

	Chair-V2		Shoe-V2	
	Acc@1	Acc@5	Acc@1	Acc@5
Vanilla Policy Gradient	61.4 %	78.6%	40.1%	71.9%
Deep Q-Learning	61.9%	78.9%	40.7%	71.8%
TRPO	60.8%	78.2%	39.8%	70.2%
PPO Actor-Only KL	62.8%	78.5%	42.3%	72.8%
PPO Actor-Only Clipped	63.9%	78.9%	43.1%	74.5%
PPO Actor-Critic KL	63.8%	78.7%	42.1%	73.7%
PPO Actor-Critic Clipped	64.8%	79.1%	43.7%	74.9%

2 Comparative Study with different reward functions

We conducted experiments with different possible reward designs as shown in Table 2. Empirically, we found that combining rewards from both ranking and feature embedding space through triplet loss offers most optimum performance.

*Interned with SketchX

Table 2: Performance analysis using different reward designs.

	Chair-V2		Shoe-V2	
Rewards	Acc@1	Acc@5	Acc@1	Acc@5
-rank	63.5 %	78.6%	42.6%	73.7%
$\frac{1}{rank}$	64.2%	78.8%	43.2%	74.2%
-$\mathcal{L}_{triplet}$	62.4%	78.1%	41.6%	72.7%
$\frac{1}{\mathcal{L}_{triplet}+\epsilon}$	60.2%	77.3%	38.8%	68.6%
$\frac{1}{rank} - \mathcal{L}_{triplet}$	64.8%	79.1%	43.7%	74.9%

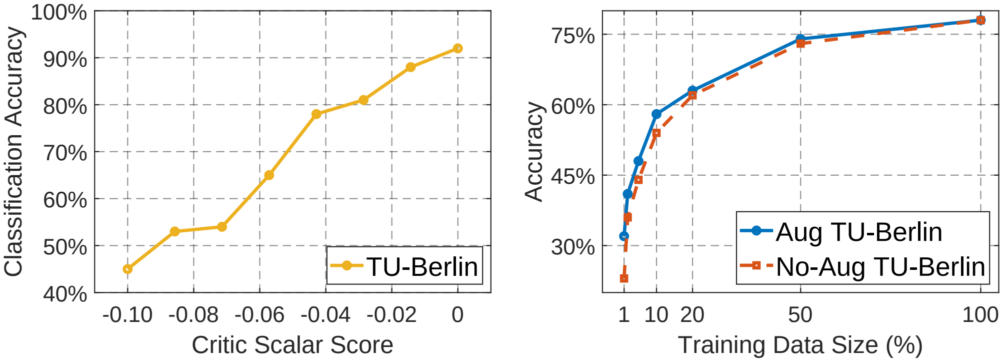

Figure 1: (a) Retrieval ability of partial sketch: correlation between critic network V(S) predicted score and ranking percentile (b) Performance at varying training data size with stroke-subset selector based data augmentation.

3 Classification Ability and Data Augmentation for sketch classification

Similar to fine-grained retrieval [4, 2], we extend our RL-based stroke-subset selector framework for classification task to judge if the critic network could be used to judge the recognition potential from partial sketch. To this end, we use negative of cross-entropy loss as the reward to train the stroke-subset selector under a pre-trained sketch classification network (Resnet50) on TU-Berlin dataset [3]. We obtain a similar correlation between critic network predicted score and classification accuracy as shown in Fig. 1. In brief, the samples having higher scalar score predicted by the critic network tends to have a higher classification accuracy, thus proving the efficiency of modelling recognition ability of sketches through our framework.

Similarly, one can use the stroke-subset selector (policy network) to augment the sketches for classification problem. Performance at varying training data regime is shown in Fig. 1 on TU-Berlin dataset.

4 Motivation on removing "fear" for sketching

By removing "fear", we meant injecting that extra confidence to the users, knowing that even if they can not sketch well, the system will still be able to return favourable results.

5 What happens with extreme cases?

The extreme case of completely random junk can be handled by our critic network, which will assign a low retrieval ability score, helping us to sidestep such unusable instances. On the other hand, critic network assigns progressively higher score for sketches from professional artists, and achieves retrieval threshold much earlier. Fig. 2 offers examples of how the critic score changes for a good/professional sketch (top) and a complete random one (bottom).

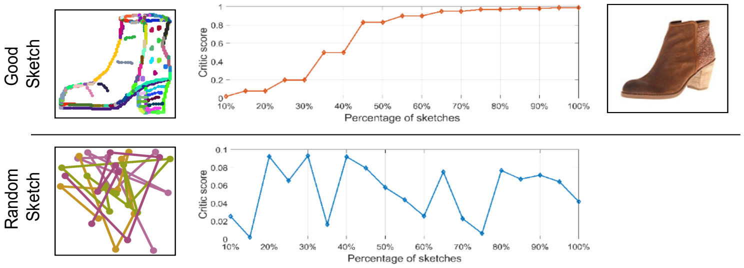

Figure 2: Critic score at progressive sketch drawing episode.

6 Clarity of binary stroke selection scheme

In our framework, we have modelled the stroke selection through categorical distribution (softmax normalisation over $\mathbb{R}^2$). However, as the reviewer suggested, one could model using Bernoulli distribution where the stroke selector would predict a single sigmoid normalised scalar value $\mathbb{R}^1$. We tried both approaches and empirically found the use of categorical distribution to be more stable with faster convergence and quantitatively better results (by 1.41% Acc@1 on ShoeV2). We will specifically mention this in the supplementary with a thorough ablative study upon acceptance.

7 Training-time comparison with baselines

For the baselines, we *do not* augment the sketches using *all possible* stroke-subset combinations (with cost $\mathcal{O}(2^N)$). Taking into account all possible stroke subsets not only slows down the training data-pipeline, but many of these augmented sketch subsets are too coarse/incomplete to convey any useful information about the paired photo. Some initial experiments indicated model collapse due to noisy gradients raised from such overly coarse/incomplete sketch-subsets. Therefore, in order to eliminate noisy gradients in the baselines, we drop strokes at random while ensuring that the percentage of sketch vector length never falls below a certain threshold – 80% was empirically found to yield optimum performance.

To ensure fair comparisons, we also keep each model training until we find no further improvement in both the loss value and accuracy on the validation set for consecutive 20K iterations. Furthermore, under our experimental setup, the training time for all baselines as well as our method lies between 12-14 hours, ensuring a largely uniform training time for all.

8 Clarity on Training dataset

We use 6051+1800 images to train both the retrieval model and stroke-subset selector. In particular, first, we pre-train the retrieval model on raster sketches. Next, we use the sketch-vector modality of the same set of sketches to train the stroke-subset selector. It should be noted that while the retrieval model trained from raster sketches is unaware of stroke-specific importance for retrieval, the stroke-subset selector intelligently manages to eliminate the noise/inconsistent sketch strokes. Testing is done on the remaining 679+200 images which are never used in either stage of the training.

9 Comparison with soft-attention

In order to deal with partial sketches, one alternative is indeed to apply soft-spatial attention in raster-space, as used in Triplet-Attn-HOLEF [6]. Through fusing Triplet-Attn-HOLEF with our Augment baseline, we devise a new baseline Triplet-Attn-HOLEF+Augment, which is able to achieve Acc@1(Acc@5) of 34.6%(68.9%) on the ShoeV2 dataset. This is slightly better than the Augment baseline but significantly falls behind our final results. This further verifies the necessity of our stroke-subset selector to deal with the erroneous/noisy strokes that are inherent to the drawing process.

10 Limitations

Cross-dataset generalisation for the stroke-subset selector in particular is an intriguing research direction, which we intend to cover in the future. Also, replacing the non-differentiable rasterization operation (sketch-vector to sketch-image) with a differentiable approximated one would be an interesting direction to explore too. This would make the whole pipeline end-to-end differentiable, so it can backpropagate the gradient calculated from triplet loss directly onto the stroke-subset selector without needing any RL-based formulation, ultimately increasing stability and pace of training.

References

[1] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 2017. 1

[2] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021. 2

[3] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *CVPR*, 2021. 2

[4] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*, 2020. 2

[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1

[6] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *CVPR*, 2017. 4