Supplementary material:
Doodle to Search: Practical Zero-Shot Sketch-based Image Retrieval

Sounak Dey*, Pau Riba*, Anjan Dutta, Josep Lladós
Computer Vison Center, UAB, Spain
{sdey, priba, adutta, josep}@cvc.uab.cat

Yi-Zhe Song
SketchX, QMUL, London, United Kingdom
yizhe.song@qmul.ac.uk

1. Introduction

In this document we present two different sections. Section 2, analyses (i) the importance of the embedding size in our architecture; (ii) the effect of different dataset split on the result in *TUBerlin-Extended* [2]. In Section 3, we show that the qualitative results which are an elaboration of the fact, that as the domain gap and abstraction of the sketches available increases, our method tends to find a better retrieved images compared to CVAE [7]. It also verifies that the datasets available in the literature are not designed for zero-shot scenario, for instance, they are not able to capture the variability in real amateur sketches.

2. Further experimentation

In this section further experiments have been done in order to validate the parameters of the proposed architecture and the choice of the reported results in-case of *TUBerlin-Extended*.

Embedding Size Table 1 presents the results obtained changing the final embedding size. The results are presented in the *Sketchy-Extended* dataset. It is a well accepted dataset without confusing categories as in the case of TUBerlin-Extended [2]. The mAP on the full database reinforces our choice of 64 dimension as mentioned in Section 5 of the original paper. Though the mAP@200 has a very marginal improvement in case of lower dimension, we would like to keep a eye on the overall mAP as we are looking towards a large scale zero-shot image retrieval. The slight change in the mAP@200 can be credited to the non-deterministic behaviour of ranking-based metrics [4].

Leakage of class information We previously followed common practice in [7], where an off-the-shelf word embedding was used without re-training. Yet, we do fully acknowledge the need to ensure no class information is leaked during training. For that, we first re-trained word2vec from scratch without the test classes from Sketchy-Extended, and obtained 0.361 mAP, which is slightly worse than the previously reported mAP of 0.369. We further trained two alternative word embeddings from scratch – GloVe and Fasttext – and report results in Table 2. It shows GloVe being superior to word2vec and Fasttext.

*These authors contributed equally to this work.

Table 1. Comparison against different sizes of embedding of our method on *Sketchy-Extended* dataset.

Dimension	**Sketchy-Extended** [2]		
	mAP	mAP@200	P@200
128	0.3508	0.4599	0.3675
64	**0.3691**	0.4606	**0.3704**
32	0.3596	**0.4691**	0.3701

Table 2. Comparison against different sizes of embedding of our method on *Sketchy-Extended* dataset.

Dataset	**word2vec** [3]	**GloVe** [5]	**fastText** [1]
Sketchy [6]	0.369	0.401	0.331
TU-Berlin [2]	0.109	0.118	0.097

Discussion on *TUBerlin-Extended* Table 3 shows the results we obtain in five different dataset split in *TUBerlin-Extended* [2] as discussed in the paper in Section 5. This in a way shows that the different splits in this dataset can have a huge effect on the zero-shot retrieval results. If seen carefully the same method performs much better in `Split-5` as compared to that in `Split-2` and `Split-3`.

3. Extra Qualitative Results

This section introduces an extended qualitative study on the three datasets. We choose 7 queries and their corresponding top-10 retrieval images. The chosen queries are the ones which best illustrates the results on the different parts of our method. These sketch queries were selected

Figure 1. Top-10 image retrieval examples for *Sketchy-Extended* [6]. All the examples correspond to a zero-shot setting. First row provides a comparison with CVAE [7] method against our pipeline. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

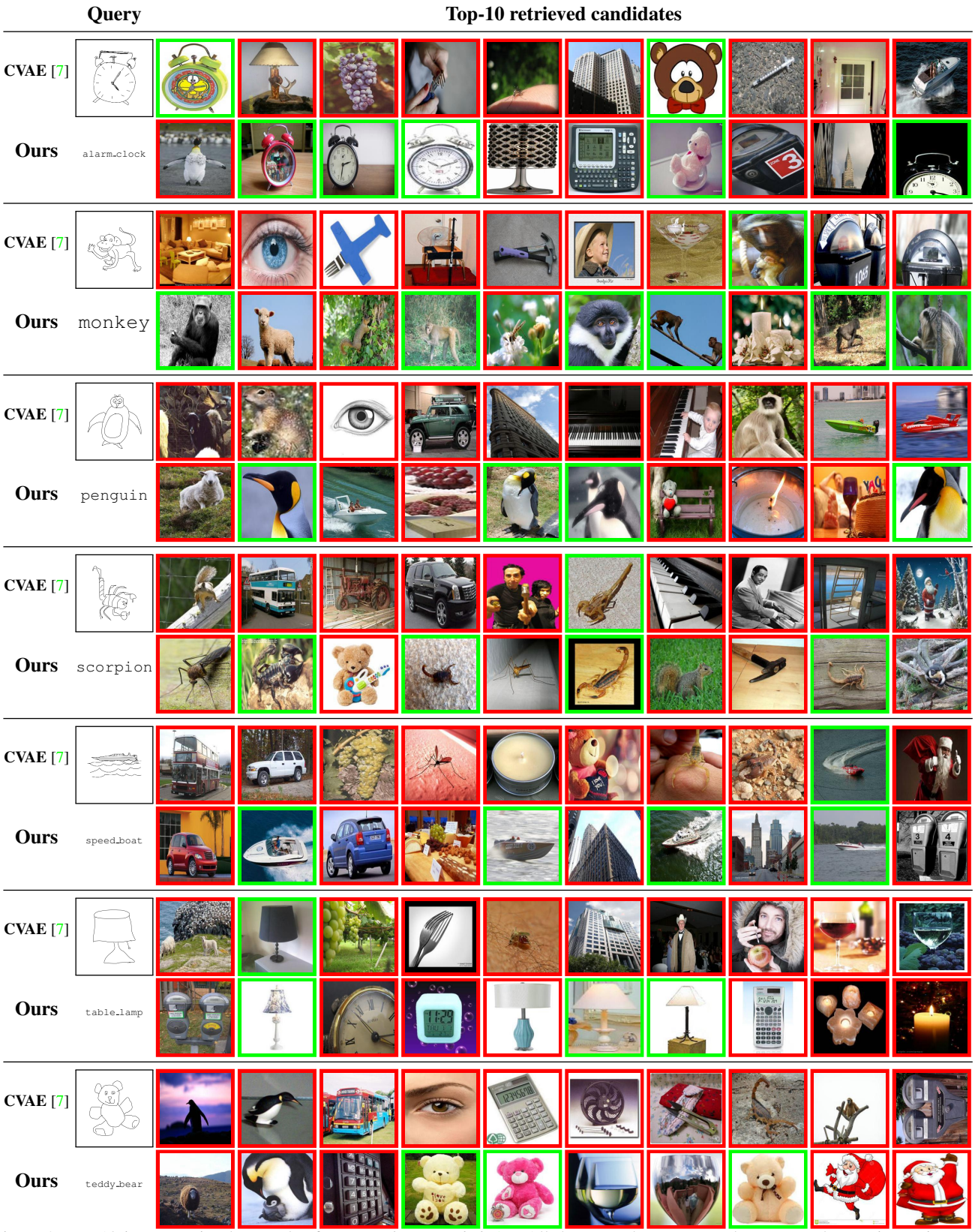

Figure 2. Top-10 image retrieval examples for *TUBerlin-Extended* [2]. All the examples correspond to a zero-shot setting. First row provides a comparison with CVAE [7] method against our pipeline. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

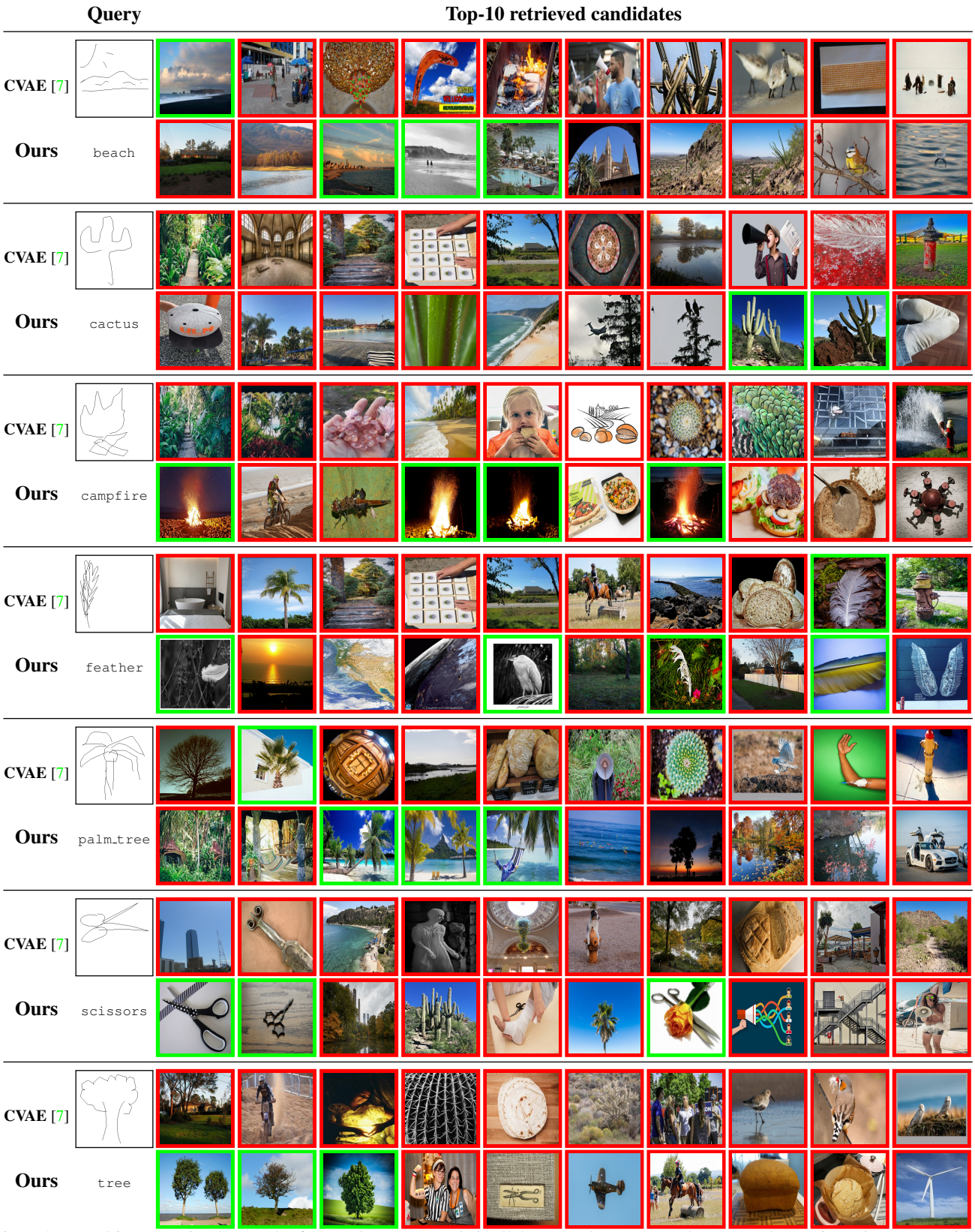

Figure 3. Top-10 image retrieval examples for *QuickDraw-Extended*. All the examples correspond to a zero-shot setting. First row provides a comparison with CVAE [7] method against our pipeline. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

Table 3. Performance on different dataset split in *TUBerlin-Extended* of our method.

Random Split	**TUBerlin-Extended** [6] mAP
Split-1	0.1184
Split-2	0.053
Split-3	0.0426
Split-4	0.1574
Split-5	0.1094
Median	**0.1094**

among approximately 12600, 2400 and 90000 test sketches in *Sketchy-Extended*, *TUBerlin-Extended* and *QuickDraw-Extended* dataset respectively.

Figure 1 presents the results for *Sketchy-Extended*. The results shows that our method performs really best due to the close correspondence of sketches and images. Like-wise the other state-of-the-art method [7] also performs good. Triplet loss helps our method to find an embedding space where the domain agnostic features are closely bounded for the same class. If observed carefully this helps to retrieve images which are visually similar to that of the sketches. In-case, of `bat` all the retrieved images have a completely spread wingspan similar to that of sketch. For `cows` and `rhinoceros` the front and the side face are really captured in the images. Even the bad retrievals in case of `mouse`, `rhinoceros` and `sword` have a lot of similar visual mappings.

Figure 3 presents the results for *TUBerlin-Extended*. Having the visual features nicely mapped in commmon embedding space we would like to demonstrate that including the semantic information ensures that the space also has some semantic correspondence. `speed_boat`, `scorpion` and `monkey` all of these sketches retrieves images that are either semantically or visual cues close to them.

Figure 3 presents the results for *QuickDraw-Extended*. Though this dataset set has the most abstractions and domain gap, our method is still able to fetch images corresponding to either the shape or the semantic feature of the queried sketch. `beach`, `cactus` and `tree` where the visual features precedes the semantic information, but in `palm_tree` the semantic information seems to play a huge part. For `feather` the retrieved images has a `bird` which are correctly in the database as it has feathers. If observed carefully the last retrieved image do contain `feather` but also contains a `fire_hydrant`. In annotation this was labelled as `fire_hydrant`. Also in case of `scissors` the fifth retrieved image has `scissors` in it, though the image is annotated as `bandage`.

According to the qualitative results we can conclude that due to the short comings of the previous datasets, we need to introduce a proper dataset that can be used by the community in the sense that none of the previously proposed datasets were designed for a *zero-shot* scenario. We also provide a benchmark and some interesting findings on the two aspects that are important while retrieving images in unseen categories.

References

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *TACL*, 2017. 1

[2] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? In *SIGGRAPH*, 2012. 1, 3

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 1

[4] A. Nicolaou, S. Dey, V. Christlein, A. Maier, and D. Karatzas. Non-deterministic behavior of ranking-based metrics when evaluating embeddings. *arXiv preprint arXiv:1806.07171*, 2018. 1

[5] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 1

[6] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *SIGGRAPH*, 2016. 1, 2, 5

[7] S. Yelamarthi, S. Krishna Reddy M, A. Mishra, and A. Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 1, 2, 3, 4, 5