Supplementary material for Sketch3T: Test-Time Training for Zero-Shot SBIR

Aneeshan Sain[1,2] Ayan Kumar Bhunia[1] Vaishnav Potlapalli* 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.sain, a.bhunia, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

1. Clarity on computational overhead:

Delving into the complexity analysis of our method we explore complexity of a relevant method in this context. The Table below compares the complexity of ZS-SAKE [2] with ours. ZS-SAKE is indeed simpler to train, and faster at test-time. The extra cost is however justifiable by (i) the ability to handle style changes in addition to novel categories, (ii) we do not dictate word embedding (as per ZS-SAKE), but just a single sketch, and (iii) we surpass ZS-SAKE [2] by a rather significant 9.31% margin (relative mAP@all).

Method	Parameters	Time per Forward Pass
ZS-SAKE [2]	27.6 mil.	25.6 ms
Ours	33.8 mil.	110.4 ms

2. Clarity on auxiliary loss used:

Without the auxiliary objective, test-time training is infeasible thus dropping model performance (Table 2, Type-I in main paper). Analysing further (Type IV-VII), we found reconstructing stroke-level details optimally conditions the encoder to a *sketch*, as it is penalised on stroke-level semantics, proving its superiority in aiding the primary objective. Furthermore, learning which strokes are significant towards boosting the primary task (via η_t in Type III) is advantageous, as some strokes inherently hold more semantic meaning in a sketch than others.

3. Clarifying experiments:

Our work differs from [4] in our latent space preservation via meta-learning, and in our auxiliary task which is optimally suited to sketches. Table below compares the performance of [1, 5] adjusted for retrieval, against ours. To clarify, in both Tables 1 and 2, our method uses test-set photo reconstruction. In Table 2, all methods involving test-time training and auxiliary task have employed test-set photo adaptation (TPA) as well. Without it, accuracy dips slightly by 0.020 mAP@all on average. Below table shows our method's accuracy in that setting (**Ours w/o TPA**).

Methods	Sketchy (ext)		TU Berlin (ext)	
	mAP@all	P@200	mAP@all	P@200
B-TENT [5]	0.483	0.574	0.405	0.521
B-SHOT [1]	0.497	0.578	0.425	0.538
Ours w/o TPA	0.561	0.620	0.495	0.642
Ours	**0.575**	**0.624**	**0.507**	**0.648**

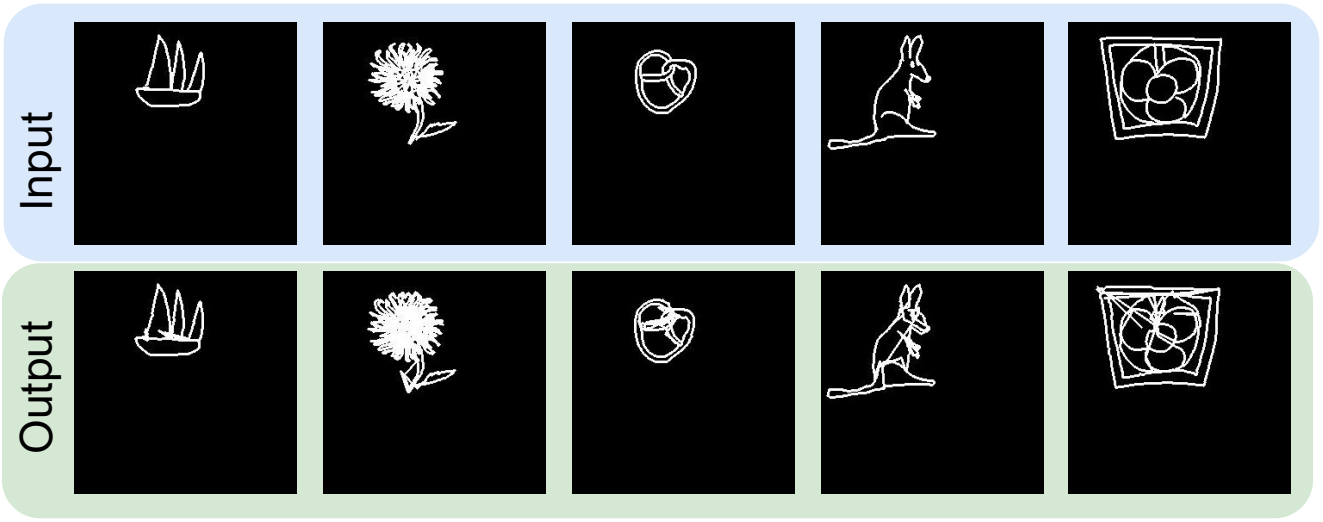
*Interned with SketchX

4. Sensitivity of hyper-parameters:

The initial estimate for some hyper-parameters like margin value of triplet loss, or initial values of inner and outer learning rates were inspired from related works [3] and optimised empirically thereafter. We have experimented by changing the ratio $\lambda_{Tri} : \lambda_{rec}$ from 7:3 to 1:1 which dipped performance to 0.510 (0.581) mAP@all (P@200) on Sketchy showing a slight sensitivity on the ratio of learning objectives. We shall include such hyperparameter sensitivity details on acceptance. For other ablation studies on sensitivity of the number of gradient steps, of both test-time training and meta-learning, or on optimal feature dimension for primary and auxiliary tasks, please refer to Fig. 4 and Fig. 5 respectively, in the main paper.

5. Additional visualisations:

Following diagram shows sketches reconstructed via the decoder (lower) against input (upper).



6. Limitations:

Despite the effective paradigm of our proposed method, there might be some cases, where the model fails to retain its learnt cross-modal knowledge of the source data. As evident from the 4^{th} sample in Figure above, the sketch reconstructed might indulge certain noisy strokes which infers that the test-time training will not always be optimal for very complex types of sketches.

References

[1] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 1

[2] 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] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 2

[4] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 1

[5] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1