Supplementary material for
Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR

Aneeshan Sain[1,2] Ayan Kumar Bhunia[1] Subhadeep Koley[1,2] Pinaki Nath Chowdhury[1,2]
Soumitri Chattopadhyay* 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

A. Alternative to Triplet Loss

Factually, for photos, intra-modal triplet loss is a self-supervised objective. For a cross-modal problem such as ours however, triplet loss can offer some leeway in better conditioning the joint photo-sketch embedding [81,70,9]. Empirically, when compared with contrastive loss [14] as a self-supervised objective (while keeping everything else the same), triplet loss performs better (45.68% on ShoeV2) further justifying our case.

B. On the need for Knowledge Distillation

During student-training, three objectives are learnt – cross-modal separation, intra-modal separation between photos and that between sketches. As learning everything together is difficult, we decouple the process by first training a teacher completely with intra-modal triplet loss on photos, and then use the trained teacher's photo discrimination knowledge to better guide the student (FG-SBIR) during its training.

C. On using additional datasets

TU-Berlin and QuickDraw are sketch-only datasets designed towards sketch classification. Some [18,23] did augment them for *category-level* SBIR [10,26], by sourcing unpaired photos. These however do not work for our instance-level setting - we need instance-level sketch-photo correspondences. The idea of abstraction-influence is very interesting, which shall be considered as a future work.

D. Dealing with scarcity of sketch-data

Distilling from unlabelled photos is beneficial as they are abundantly available, unlike sketches that require time and human effort to collect [4]. On distilling from only sketches there is minimal increment from teacher supervision (44.51% vs. 44.18 % on ShoeV2) as compared to that from photos (48.35% vs. 44.18% on ShoeV2). Faithful sketch-generation in photo-to-sketch generation tasks is challenging; it's difficult to quantify its generation-quality, and they hardly generalise to human-sketches [4]. Using CLIPasso [B] for sketch-generation instead of teacher-supervision, hence delivers a poor result of 38.57% on ShoeV2. Although works have explored augmenting sketches via stroke-dropping/deformations [81], or as line-drawings [C], resulting sketches mostly follow edge-maps thus being less reliable. On using [C] instead of teacher-supervision, we obtained a poorer result of 39.23% compared to our 48.35% on ShoeV2, thus proving our method to be simpler and more efficient.

E. Clarity on training teacher

The teacher comprises an ImageNet pre-trained PVT backbone trained on 60,502 additional photos from Sketchy (ext) [41] for Sketchy; 50,025 photos of UT-Zap50k [80] for ShoeV2 [81]; and 7,800 photos from websites like IKEA, etc [50] for ChairV2 [61].

*Interned with SketchX

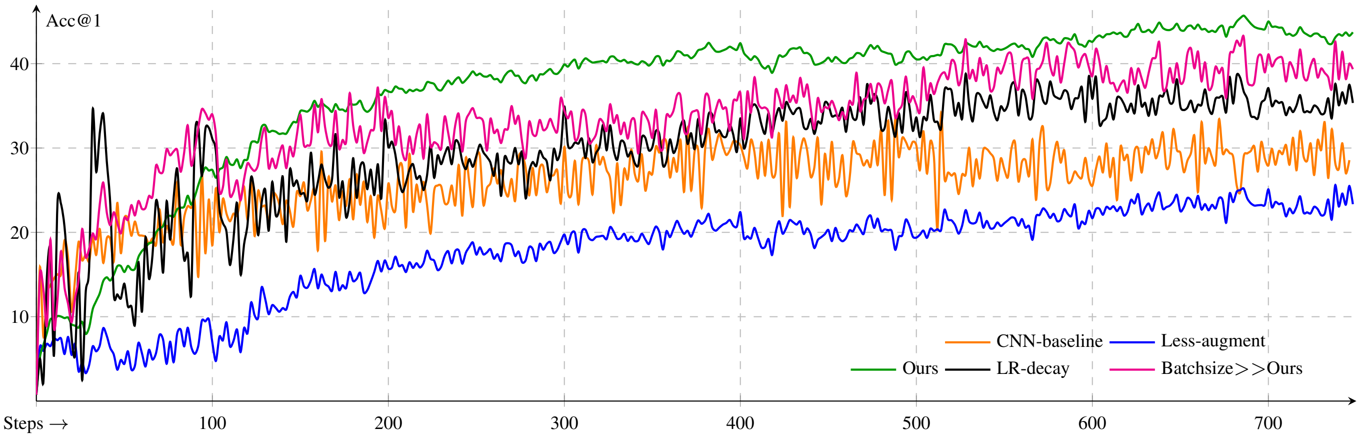
F. Optimisation for multi-task objectives

We used the available toolbox of *WandB Sweeps* for quick tuning of hyper-parameters, which provided 48.35% Acc@1 on ShoeV2. Even on using complex loss balancing approach of [D], we obtain a close 47.94%. Furthermore, changing the hyper-parameter values by ±10%, causes a mere ±0.5% change in Acc@1 on ShoeV2 [81]. This proves that our method despite needing five loss objectives, is quick to tune, thus being easily reproducible.

G. Clarity on training stability

- Learning rate decay: We used exponential rate decay with initial learning rate of 0.001 and decay factor of 0.2.
- Large batch-size: We used 256 batch-size via gradient accumulation [A] on 4 V100-GPU machines.
- Reducing augmentations: We used augmentation (random horizontal flipping only) on just 30% of training data.

Plots below show the above methods' implementation on top of CNN-Baseline. Least gittering in *Ours* shows our EMA approach to be superior.



Evaluation accuracy at every 100th training-step [Best if zoomed].

H. Further clarity on experimental results

As we intended to show PVT [74] is a better backbone than the earlier CNN-based ones, we compared prior state-of-arts to our method using different backbones. Furthermore, for the methods having code available, we replaced their backbones with PVT, only to obtain inferior results (32.68% for [81] and 34.12% for [70]), thus proving ours as better. Furthermore, our method surpasses by 8.33% on ShoeV2, against a contemporary method of TC-Net [E], despite having a lesser complexity of training and simpler loss objectives than the latter.

References:

[A] Yash Patel et. al. "Recall@k Surrogate Loss with Large Batches and Similarity Mixup" CVPR'22.

[B] Yael Vinker et. al. "CLIPasso: Semantically-Aware Object Sketching" SIGGRAPH'22.

[C] Caroline Chan et. al. "Learning to generate line drawings that convey geometry and semantics" CVPR'22.

[D] Rick Groenendijk et. al. "Multi-Loss Weighting with Coefficient of Variations" WACV'21.

[E] Hangyu Lin "Tc-net for isbir: Triplet classification network for instance-level sketch based image retrieval" WACV'21.