Supplementary material:Photo Pre-Training, But for Sketch

Ke Li[1,2] Kaiyue Pang[2] Yi-Zhe Song[2]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
[2] SketchX, CVSSP, University of Surrey, United Kingdom

Experiment Settings

We validate our method on all five publicly available product-level FG-SBIR datasets, namely QMUL-Shoe-V1, QMUL-Shoe-V2, QMUL-Chair-V1, QMUL-Chair-V2 and QMUL-Handbag and follow their original train/test split for fairness [1]. Existing FG-SBIR works have employed various choices as their network backbones, with some common choices such as Sketch-a-Net [15, 16, 18], VGG-16 [12], DenseNet-169 [7], Inceptionv1 [11], InceptionV3 [2–4, 13, 14, 19]. Most of these backbones have been pre-trained on the ImageNet benchmark [17] with 1000-way classification (*i.e.* loaded from an off-the-shelf model zoo like `torchvision.models`), with few choosing a more tailored pre-training approach including learning on a third-party task and dataset [11, 18]. We choose ResNet50 [5] as our network backbone due to its increasing popularity and superiority as a better ConvNet architecture [8], while being still on a roughly similar scale (if not smaller) with those of baseline models.

We follow the Siamese choice [10] where sketch and photo representation learning share same set of parameters. All input data are first resized to 299 x 299 and randomly cropped to 256 x 256. Another data augmentation strategy used is horizontal flipping. We set the batch size to 16 and train 600 epochs for all task settings with a SGD optimiser and momentum value of 0.9[1]. We report top 1 ranking performance ($acc@1$). To obtain the neighbourhood matrix R, we extract the feature from the last conv layer for each photos in the training set and form all possible photo triplets before ranking their relative distance. We pre-compute R once before conducting any online FG-SBIR learning. For both multi-task and meta learning setting: the learning rate is set to 1e-3 (η_s, η_t) with Δ_{sp} = 0.1 and Δ_{NT} = 0.01. Ratio β/α = 1 if not otherwise mentioned. A pytorch style algorithmic schematics can be found as follows.

[1] While Adam optimiser [6] is commonly adopted by existing FG-SBIR works and occasionally leads to a better peak performance, we find such superiority in reported numbers with trade off from worse reproducibliity troubling. We therefore choose SGD-M which yields a more stable test-time performance. We also experience with other more advanced optimisers like Adam-W [9] but do not observe extra gains.

ALGORITHM 1
Pseudo code for implementing Eq. 9 of the main text, which usually takes a few more lines in practice (**line 6,8,10**).

1: **Input:** $\{s, p^+, p^-\} \in \mathbb{R}_{B\times H\times W\times C}$ – triplet batch inputs. $R \in \mathbb{R}_{B\times B\times B}$ – relative neighbourhood ranking matrix. K – number of local neighbourhood relations simulated per update. θ_n, $\{\eta_s, \eta_t, \Delta_{\mathrm{FG-SBIR}}, \Delta_{\mathrm{NT}}, \epsilon\}$ – current model state and some hyperparameters.
2: **Output:** θ_{n+1} – updated model state directed by both $L_{\mathrm{FG-SBIR}}$ and L_{NT}

3: *% calculate gradient of θ_n w.r.t. $L_{\mathrm{FG-SBIR}}$*
4: $\hat{g}$:= θ_n.grad(max(MSELoss($\Psi(s, \theta_n)$, $\Psi(p^+, \theta_n)$) - MSELoss($\Psi(s, \theta_n)$, $\Psi(p^-, \theta_n)$) + Δ_{sp}, 0))
5: *% calculate intermediate model state θ_{temp} .*
6: θ_{temp} := θ_n - $\eta_s\hat{g}$
7: *% form K random batch-wise triplets for each s.*
8: $\{p_*^+, p_*^-\}$:= random.sample(unique(meshgrid(range[1, B], range[1, B])), K)
9: *% calculate gradient of θ_{temp} w.r.t. L_{NT}*
10: $\bar{g}$:= θ_{temp}.grad(max(**R(:, [s, p_*^+, p_*^-])**×(MSELoss($\Psi(s, \theta_{temp})$, $\Psi(p_*^+, \theta_{temp})$) - MSELoss($\Psi(s, \theta_{temp})$, $\Psi(p_*^-, \theta_{temp})$)) + Δ_{NT}, 0))
11: % writing down the final update rule
12: θ_{n+1} := θ_n - $\eta_s\hat{g}$ - $\eta_t\bar{g}$

References

[1] SketchX!-Shoe/Chair Fine-grained-SBIR dataset. http://sketchx.ai, 2022. 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. 1

[3] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *CVPR*, 2022. 1

[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. 1

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

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. Tc-net for isbir: Triplet classification network for instance-level sketch based image retrieval. In *ACM MM*, 2019. 1

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 1

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[10] Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017. 1

[11] Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020. 1

[12] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *ECCV*, 2018. 1

[13] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In *BMVC*, 2020. 1

[14] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 1

[15] Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Xiang Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *BMVC*, 2016. 1

[16] 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 *ICCV*, 2017. 1

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[18] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1

[19] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Fine-grained instance-level sketch-based image retrieval. *International Journal of Computer Vision*, 2021. 1