Supplementary Material for Sketch Down the FLOPs: Towards Efficient Networks for Human Sketch

Aneeshan Sain[1] Subhajit Maity[2*] Pinaki Nath Chowdhury[1] Shubhadeep Koley[1]
Ayan Kumar Bhunia[1] Yi-Zhe Song[1]
[1] SketchX, CVSSP, University of Surrey, United Kingdom.
[2] Department of Computer Science, University of Central Florida, United States.
{a.sain, p.chowdhury, s.koley, a.bhunia, y.song}@surrey.ac.uk; Subhajit@ucf.edu

Further analyses:

Table A. Performance & compute for Transformer-based networks

Model (on 224×224)	Swin-B	PVT-L	ViT-B/16	DeiT-B/16	CvT-21
FLOPs (approx.)	15.4G	9.8G	17.8G	17.6G	9.2G
Params (approx.)	88M	61.4M	86M	87M	32M
Top-1 (ShoeV2)	40.71%	44.18%	16.28%	35.62%	41.58%

Table B. Results on transformer-based architectures.

Model	Self (%)	SketchyNetV1 (%)	SketchyNetV2 (%)
SketchPVT [5]	44.18	43.01	42.47
Swin-B [5]	40.71	38.94	38.02
EfficientViT-M5 [4]	38.76	37.29	36.58
SketchAbstract [3]	45.31	44.06	43.59
Partial [6] (full-sketch)	32.87	31.33	30.79
ViT [5] / MobileNetV2	16.28	12.53	11.64
ViT [5] / MobileFormer	16.28	14.86	13.65

Generalization to transformers:

Table A shows the Top-1 score on ShoeV2 of popular transformers (trained like [5]), and also experiment (Top-1 on ShoeV2) using suggested methods (from [5]) as teachers and MobileNetV2 as student. To recap, the aim is to *retain the accuracy of the teacher model* in its smaller variants of SketchyNetV1 and SketchyNetV2. Two more experiments use the same teacher of pretrained Vit (trained on Triplet Loss), to train *(a)* MobileFormer [1] (transformer) and *(b)* MobileNetV2 (CNN). Tabulated results (Table B) show SketchyNetV2 variants to hold nearly similar accuracy as their teachers, akin to our findings in Table 2, thus validating our method.

Other ablations and experiments: *(i)* While [6] uses deep reinforced attention regression to tackle unnecessary strokes for retrieval from *partial* sketches and [5] explores transformers for accuracy our goal is to deliver the *same* accuracy of a large model at a much *reduced compute* thus addressing the *sparsity* in a sketch. *(ii)* On varying λ as (0.1, 0.15, $\cdots$ 0.9), accuracy falls when $\lambda > 0.7$ or $\lambda < 0.45$, residing optimally at $\lambda = 0.5$, thus suggesting an equal impact of triplet loss and KD component.

Distinguishing our work from [5] for clarity: While our work might share certain components with [5], it fundamentally differs in motivation, methodology, and novelties: *(i)* ours is the **first** investigation into efficient sketch networks adapted from photo-based ones. *(ii)* Unlike [5], which applies KD on fixed-resolution sketches, our novel abstraction-aware canvas-size selector dynamically optimizes sketch resolutions, achieving a 97.92% reduction in FLOPs while retaining accuracy. *(iii)* Our KD framework uniquely uses sketch features from the teacher, always extracted at *full* resolution, to guide the student operating on *varying* resolutions, invoking **scale invariance**, a property ignored in [5], and crucial for our scenario. *(iv)* We specifically address the sparsity and abstraction inherent to sketches, while [5] focuses primarily on dense visual representations.

Miscellaneous: *(i)* Compared to MSE or MAE losses, which are prone to noisy outliers, **Huber loss** provides robustness by balancing convergence and stability [2]. It enhances generalization, reducing sensitivity to small deviations and large outliers, particularly benefiting our cross-modal FGSBIR task. *(ii)* Interestingly, more sketches were downsized for ShoeV2 (22.31% at 128x128, 38.12% to 64x64, 17.64% to 32x32) compared to FSCOCO (16.75% at 128x128, 1.2% to 64x64) dataset, likely because the latter holds much more detailed sketches.

*Work done as an intern at SketchX before joining UCF.

References

[1] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *CVPR*, 2022. 1
[2] Shouyou Huang and Qiang Wu. Robust pairwise learning with huber loss. *Journal of Complexity*, 2021. 1
[3] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. How to handle sketch-abstraction in sketch-based image retrieval? In *CVPR*, 2024. 1
[4] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *CVPR*, 2023. 1
[5] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR. In *CVPR*, 2023. 1
[6] Dingrong Wang, Hitesh Sapkota, Xumin Liu, and Qi Yu. Deep reinforced attention regression for partial sketch based image retrieval. In *ICDM*, 2021. 1