Supplementary material for Sketch2Saliency: Learning to Detect Salient Objects from Human Drawings

Ayan Kumar Bhunia[1] Subhadeep Koley[1,2] Amandeep Kumar* Aneeshan Sain[1,2]
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.bhunia, s.koley, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

A. Sketch Vector Normalisation

Firstly, denoting sketch-vectors as a sequence of five-element vectors with *off-set* values over absolute coordinate is common in sketch/handwriting literature [1, 4, 7]. It mainly models the free-flow nature of drawing via GMM [6]. Regressing to absolute coordinates otherwise, results in mean output [2] without any instance-specific variation. Secondly, offset makes sketch invariant to drawing-position in a sketch-canvas (Fig. 2). Keeping rest of the design same, replacing GMM-based loss by standard l_1 loss based absolute coordinate regression, reduces max F_β value on ECSSD dataset to 0.652 from 0.781 (ours), thus validating the need of off-set and GMM-based design. Furthermore, sketch-vector-length varies across samples in a batch, a specific pen-state for end-of-drawing is needed to mask out loss computation from zero-appended tails of sketch-vector.

B. Multi-scale 2D Attention Module

(i) J is an intermediate tensor, which is aware of three factors: (a) local and (b) neighbourhood information, of $\mathcal{B}$, and (c) previous state of auto-regressive decoder for sequential modelling. Later, $J \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times d}$ helps compute the attention-map α. *(ii)* $\mathcal{B}(i,j)$ signifies the 1×1 convolution applied at every (i,j) spatial position for local-information modelling. *(iii)* The first two terms are tensors of size $\mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times d}$ and $W_s s_{t-1}$ is a vector of size $\mathbb{R}^d$ which is broadcasted (standard PyTorch convention) to the required spatial size for addition. *(iv)* Eq. 1 is employed using 1×1 convolution with kernel W_a, where softmax is applied across the spatial size, $g_t \in \mathbb{R}^d$. *(v)* Output of the last three max-pooling layers of VGG-16 are used for multi-scale feature aggregation which have a spatial down-scaling factor of $8, 16, 32$, respectively. *(vi)* During single-scale ablative setup, we only use the output feature-map of the last pooling layer $\mathcal{F}^l$.

C. Scribbles vs Sketch

Despite taking more time, sketches hold way more structural *and* semantic cues than the much *sparser* and *zero-semantics* scribbles [9]. Also, temporal aspect of sketches may initiate future works on *relative saliency* of objects at scene level.

D. Advantage of Sequential Stroke Modelling

Firstly, as free-hand sketches are not edge-aligned with its paired photo, there is no direct way to post-process the sketch-coordinate to get aligned key-points attending the silhouettes/corners of the object. However, following [8], we design a baseline as: Apply a spatial attention module on backbone-extracted feature-map followed by global-average pooling and a fully-connected layer to directly predict the $T = 100$ (fixed via RDP algorithm) absolute coordinates, assuming that the network will attend salient regions via spatial attention. Here max F_β on ECSSD dataset falls to 0.692. Importantly, we can not use off-set based representation here as by definition it relies on sequential modelling explicitly. Therefore, we adopted sequential stroke modelling with 'look-back mechanism' via multi-scale 2D-attention for our saliency framework.

(i) Moreover, our 1-layer LSTM based auto-regressive network is quite standard for works [1,3,4,7] like image captioning, handwriting/sketch generation, text recognition, and simple to optimise. *(ii)* Removing pen-state prediction hurts accuracy

*Interned with SketchX

(max F_β: 0.741) as the model gets confused for large-jumps at stroke-transitions via off-set-based modelling. *(iii)* Furthermore, this temporal aspect of sketch may potentially convey relative saliency (sorted by stroke order) – an interesting topic for future study.

E. Correlating stroke-prediction and saliency-map quality

We used photo-to-sketch generation as an auxiliary task to solve the *saliency detection* problem, where performance of the *latter is crucial* but *not the former*. Moreover we found that, while avg. $F_\beta^{\max}$ for samples whose sketch-generation metric, *log-loss*, (mean of Eq. 6 in nats [5] is lower (better) than -1100, comes to 0.824, the same for samples with *log-loss* higher (worse) than -1000 drops to 0.736, thus justifying the correlation quantitatively.

References

[1] 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

[2] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: A competitive sketching ai agent. so you think you can beat me? In *Siggraph Asia*, 2020. 1

[3] Ayan Kumar Bhunia, Aneeshan Sain, Amandeep Kumar, Shuvozit Ghose, Pinaki Nath Chowdhury, and Yi-Zhe Song. Joint visual semantic reasoning: Multi-stage decoder for text recognition. In *ICCV*, 2021. 1

[4] Nan Cao, Xin Yan, Yang Shi, and Chaoran Chen. Ai-sketcher: a deep generative model for producing high-quality sketches. In *AAAI*, 2019. 1

[5] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 2

[6] David Ha and Douglas Eck. A neural representation of sketch drawings. *ICLR*, 2017. 1

[7] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *CVPR*, 2018. 1

[8] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *CVPR*, 2019. 1

[9] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020. 1