Recursive Contour-Saliency Blending Network for Accurate Salient Object Detection
(Supplimentary Material)

Yun Yi Ke
Computer Vision & AI Technology Lab
Open8 Singapore
yunyikeyyk@gmail.com

Takahiro Tsubono†
Computer Vision & AI Technology Lab
Open8 Singapore
tsubonot@open8.com

1. Content

In this supplementary file, we provide more details of our proposed network, RCSBNet. Specifically,

- in Section 2, we present more details and analysis of our model.
- in Section 3, we present more comparisons of saliency predictions between RCSBNet and other state-of-the-art models.
- in Section 4, we provide more comparisons of contour predictions between RCSBNet and state-of-the-art models using contour information, which are ITSD [6] and PoolNet [2].

2. Experimental Results and Model Analysis

Predictions of Intermediate Layers. In the figure below, we illustrate how final predictions are generated stage by stage in our RCSBNet.

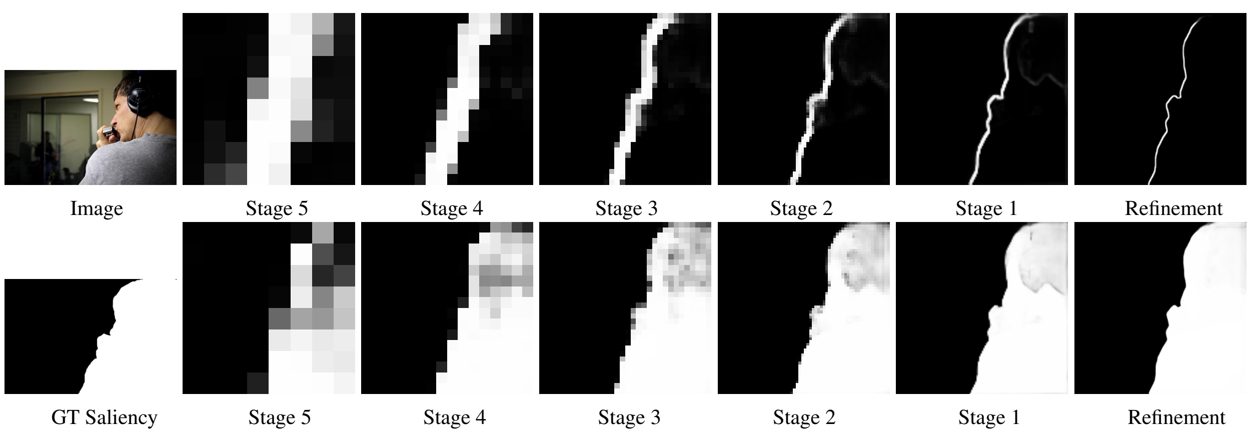

Figure 1: Visualizations of intermediate predictions from each decoder stage and refinement module. First row: input image and contour predictions from 5 decoder stages and refinement module. Second row: ground truth saliency and saliency predictions from 5 decoder stages and refinement module. Note that prediction accuracy is improved from stage 5 to stage 1 due to the training against accuracy-related loss. The confidence of prediction gets enhanced at the refinement stage due to the training against confidence loss.

†Corresponding author.

For 5 decoder stages and refinement module, the output shapes are 8×8, 16×16, 32×32, 64×64, 128×128, and 256×256, respectively. Predictions are supervised against accuracy-related loss in stages 1 to 5 and confidence loss in the refinement module. As shown in Fig. 1, saliency prediction in stage 5 contains false negatives but gets enhanced in stage 1. Meanwhile, the confidence of saliency prediction in stage 1 gets improved after the refinement module.

Using Confidence Score as an Evaluation Metric. As mentioned in Sec. 3.6 in the paper, we introduced a confidence score, W_c, for each pixel $x_{i,j}$ in prediction: $W_c = \beta * x_{i,j} * (1 - x_{i,j})$, where β is empirically set to 2. The score will be 0 if the prediction is binary and reach the maximum value if $x_{i,j} = 0.5$. Thus this score can also be used as an evaluation metric to measure how close is the saliency prediction against the binary ground truth. We define the average confidence score, C_β, among all images in a dataset as:

$$C_\beta = \frac{1}{n \times p \times q} \sum_{k=1}^{n} \sum_{i=1}^{p} \sum_{j=1}^{q} \beta * x_{i,j} * (1 - x_{i,j}) \tag{1}$$

where n represents total images in the dataset and p, q stand for image dimension. Thus, in addition to the quantitative comparison listed in Section 4.4 Table 1 in the paper, we also compare our model under average confidence score $C_{\beta=2}$ with 7 state-of-the-art methods in the table below.

Table 1: Quantitative comparisons between RCSBNet and other 6 methods on five benchmark datasets in terms of the $C_{\beta=2}$. **Red**, **Green**, and **Blue** indicate the best, second best and third best performance. Subscripts stand for year of the paper.

Method	Contour Information	DUTS-TE	DUT-OMRON	PASCAL-S	ECSSD	HKU-IS
		$C_{\beta=2} \downarrow$	$C_{\beta=2} \downarrow$	$C_{\beta=2} \downarrow$	$C_{\beta=2} \downarrow$	$C_{\beta=2} \downarrow$
F3Net$_{20}$ [4]	✗	**.0123**	**.0143**	**.0146**	**.0127**	.0109
MINet$_{20}$ [3]	✗	**.0143**	**.0160**	**.0190**	**.0150**	.0139
GCPA$_{20}$ [1]	✗	.0196	.0196	.0220	.0211	.0203
EGNet$_{20}$ [5]	✓	.0191	.0206	.0217	.0199	**.0059**
PoolNet$_{20}$ [2]	✓	.0182	.0195	.0209	.0193	**.0057**
ITSD$_{20}$ [6]	✓	.0166	.0199	.0196	.0166	.0154
Ours	✓	**.0085**	**.0102**	**.0106**	**.0086**	**.0081**

Effectiveness of the number of recursions. To investigate the effectiveness of recursion R, we gradually increase the recursion from 1 to 4 and measure $\overline{F_\beta}$, MAE, E_ξ, and $F_\beta^\omega \uparrow$ accordingly on DUTS-TE and ECSSD datasets. As shown in Table 2, when R equals 3, the model yields the best performance.

Table 2: Ablation study for the effect of recursion number. When R=3, the best results are obtained.

	DUTS-TE				ECSSD			
	$\overline{F_\beta} \uparrow$	$M \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$\overline{F_\beta} \uparrow$	$M \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$
R=1	.836	.037	.899	.821	.917	.038	.917	.901
R=2	.844	.036	.901	.830	.925	.035	.918	.908
R=3	**.855**	**.034**	**.903**	**.840**	**.927**	**.033**	**.923**	**.916**
R=4	.850	.038	.900	.832	.922	.036	.916	.910

3. More Visual Comparisons on Saliency Predictions

We list more images in Fig. 2 for visual comparisons on saliency predictions. It is well demonstrated that our proposed RCSBNet can consistently generate accurate and complete saliency predictions compared with other state-of-the-art models. Besides, our model can detect small salient objects while predictions from other methods contain either incomplete predictions or a considerable amount of false positives.

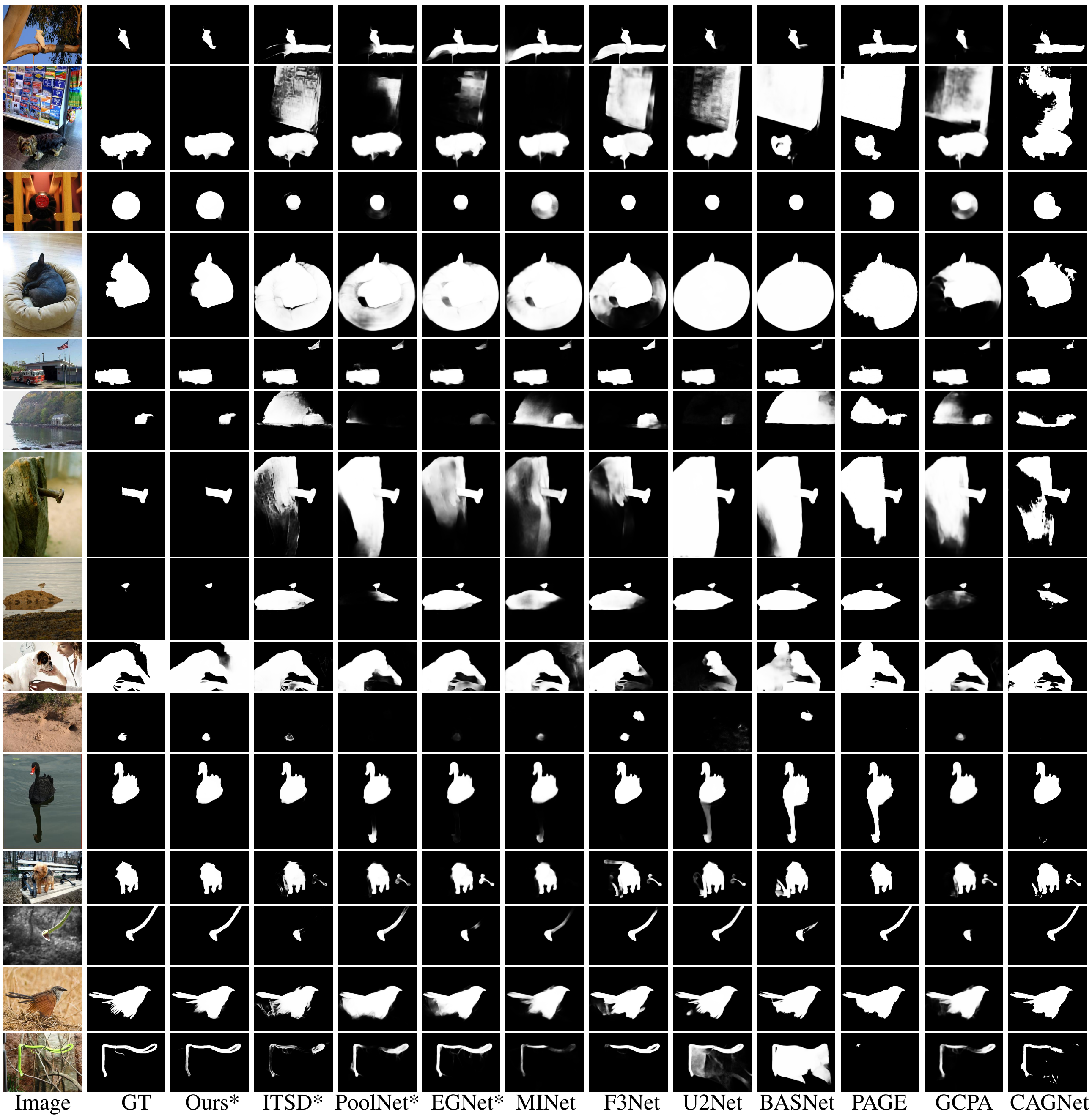

Figure 2: Visual comparison of salient object predictions between our method and 10 state-of-the-art networks. * stands for models utilizing contour information.

4. Visual Comparisons on Contour Predictions

We list more images in Fig. 3 for visual comparisons on contour predictions between our RCSBNet, ITSD, and PoolNet. As illustrated below, RCSB can generate more complete and better contour predictions. This is due to the stage-wise feature extraction (SFE) module and the effectiveness of the recursive mechanism where contour and saliency are blended multiple times.

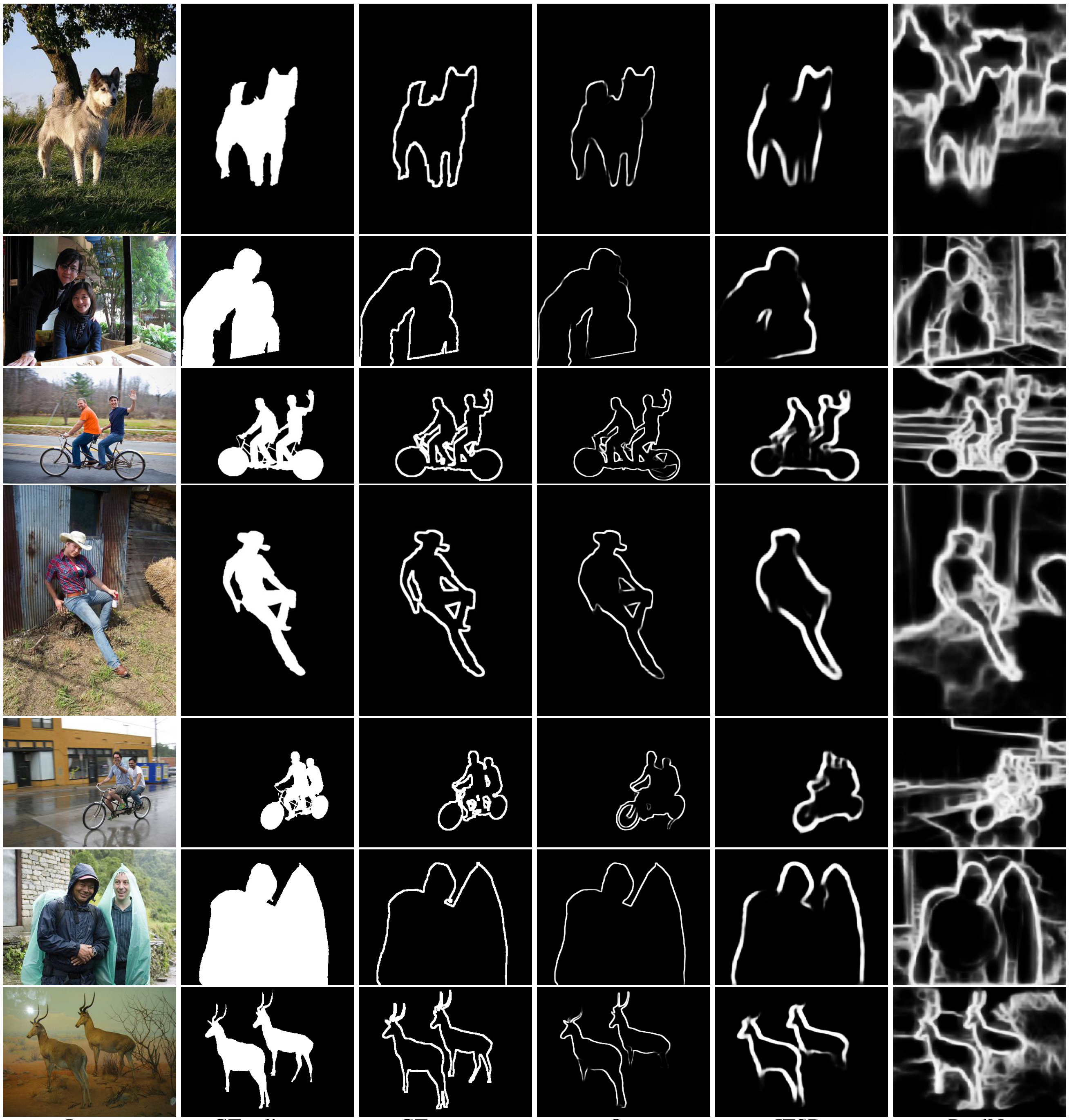

Image GT saliency GT contour Ours ITSD PoolNet

Figure 3: Visual comparison of contour predictions between our method, ITSD and PoolNet. Ground truth contours are obtained via erosion and dilation with kernel size of 5.

References

[1] Z. Chen, Q. Xu, R. Cong, and Q. Huang. Global context-aware progressive aggregation network for salient object detection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10599–10606. AAAI Press, 2020.

[2] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang. A simple pooling-based design for real-time salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3917–3926. Computer Vision Foundation / IEEE, 2019.

[3] Y. Pang, X. Zhao, L. Zhang, and H. Lu. Multi-scale interactive network for salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9410–9419. IEEE, 2020.

[4] J. Wei, S. Wang, and Q. Huang. F3net: Fusion, feedback and focus for salient object detection. *CoRR*, abs/1911.11445, 2019.

[5] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng. Egnet: Edge guidance network for salient object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8778–8787. IEEE, 2019.

[6] H. Zhou, X. Xie, J. Lai, Z. Chen, and L. Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9138–9147. IEEE, 2020.