# Supplemental Material for 'Just a Hint: Point-Supervised Camouflaged Object Detection'

Huafeng Chen[1,2], Dian Shao[1,2] ✉, Guangqian Guo[1,2], and Shan Gao[1,2] ✉

[1] Unmanned System Research Institute, Northwestern Polytechnical University
[2] National Key Laboratory of Unmanned Aerial Vehicle Technology
{chf,guogq21}@mail.nwpu.edu.cn, {shaodian,gaoshan}@nwpu.edu.cn

## 1 Framework Details.

### 1.1 Former Solutions Fails in COD.

Existing point-based salient object detection (SOD) method PSOD [7] uses edge maps to generate proposal regions. Other two-stage methods such as SPOL [22], use the first stage to generate a prediction map as the proposal region. But, the above methods are not appropriate for PCOD, because camouflaged objects and backgrounds have very low contrast and ambiguous edges, which leads to poor edge maps or prediction maps using only one point annotation, as shown in Figure 1.

### 1.2 Detailed Structure of the Encoder.

As shown in Figure 2, we use PVT [21] as the backbone and put an input image $I \in \mathbb{R}^{3\times H\times W}$ into the backbone to get the output features $Feat_i$ for the $i$-th. Then, we get the multi-scale features $(F_{eat1}, F_{eat2}, F_{eat3}, F_{eat4})$ with $(\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32})$ resolution of input images. We downsize the channel dimension of $F_{eati}$ into 64 by using $3 \times 3$ convolutional layers. Next, these feature maps are unified into the same size by an up-sampling operation, and combined through the concatenation. Finally, the output map $\hat{S}\in \mathbb{R}^{1\times W\times H}$ is obtained by the $3\times 3$ convolution layer.

### 1.3 The Structure of $g$ in Representation Optimizer.

In Figure 3, we show the structure of the small network $g$ in the representation optimizer. Specifically, the predictor consists of an MLP with 2 layers. The dimensions of the input and output are equal to the number of image pixels, while the dimension of the hidden layer is one-quarter of the input dimension, creating a bottleneck structure for $g$.

✉ Corresponding Authors

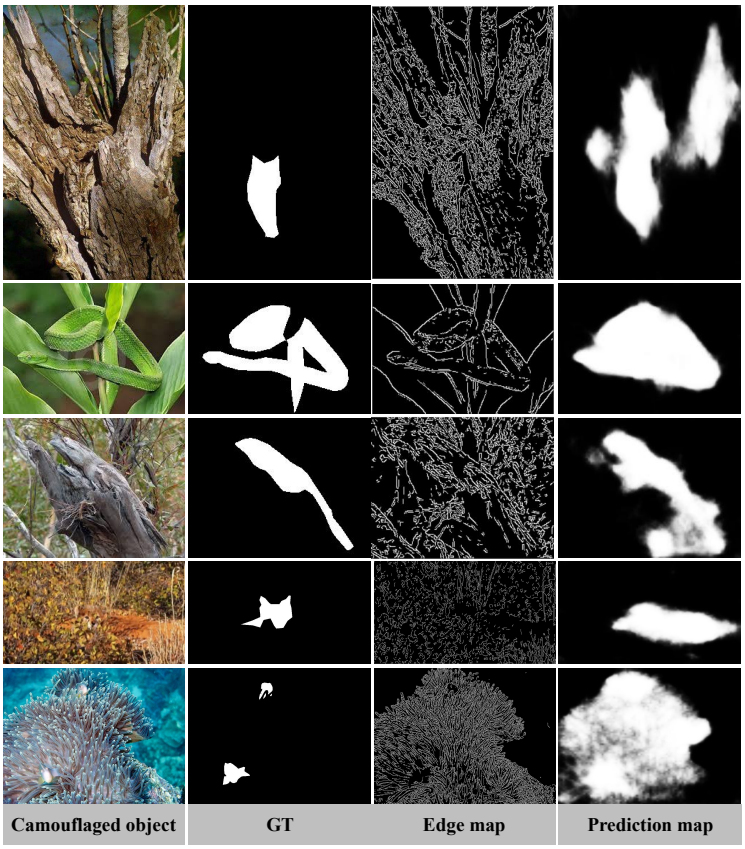


**Fig. 1:** Edge map and prediction map fail in COD

## 2 Experiment Details.

### 2.1 Experiments on Scribble Dataset.

To demonstrate strong generalization performance and further validate our model, we evaluate the proposed model on the scribble dataset. The experimental dataset, testing metrics, and implementation details remain consistent with those of the previous point-labeled dataset.

### 2.2 Experiments on Salient Object Detection.

In order to verify the generalization of the network design, we evaluate the proposed model on the SOD task.
**Datasets.** Our experiment on SOD is based on the existing four SOD datasets, ECSSD [25], DUT-O [26], HKU-IS [11], and DUTS-test [19]. We only use the

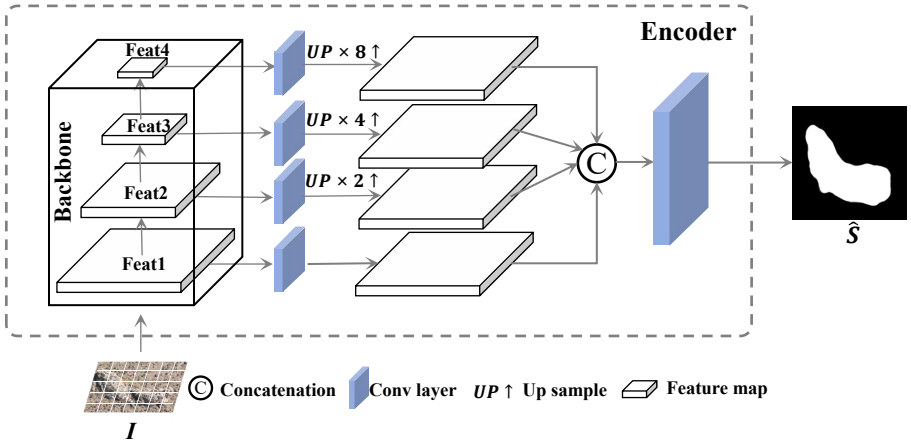


**Fig. 2:** The architecture of encoder.

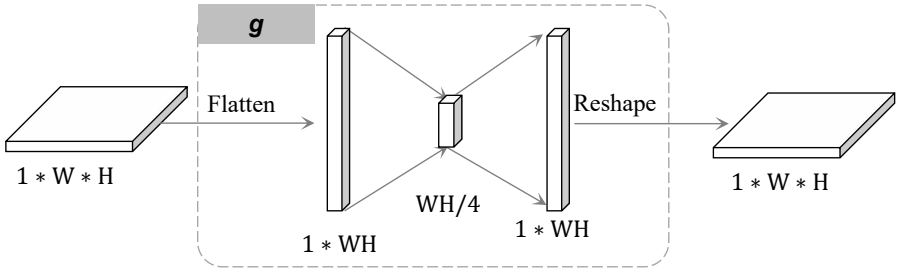


**Fig. 3:** The architecture of $g$.

training set of DUTS for training. During the test phase, we use the remaining data for inference.

**Implementation Details.** We use the stochastic gradient descent optimizer with a momentum of 0.9, a weight decay of $5e-4$, and triangle learning rate schedule with maximum learning rate $1e-3$. The batch size is 8, and the training epoch is 60. During training and inference, input images are resized to $512 \times 512$.

## 3 More Results and Analysis.

### 3.1 PCOD Dataset.

Pixel annotation is time-consuming and labor-intensive, while scribble labeling is both time-consuming and difficult to control for quality. To overcome these limitations, we propose a new dataset PCOD with point annotation. Specifically, three annotators participate in the annotation task. For each image, we randomly choose one annotation from the three to reduce individual bias. Additionally,

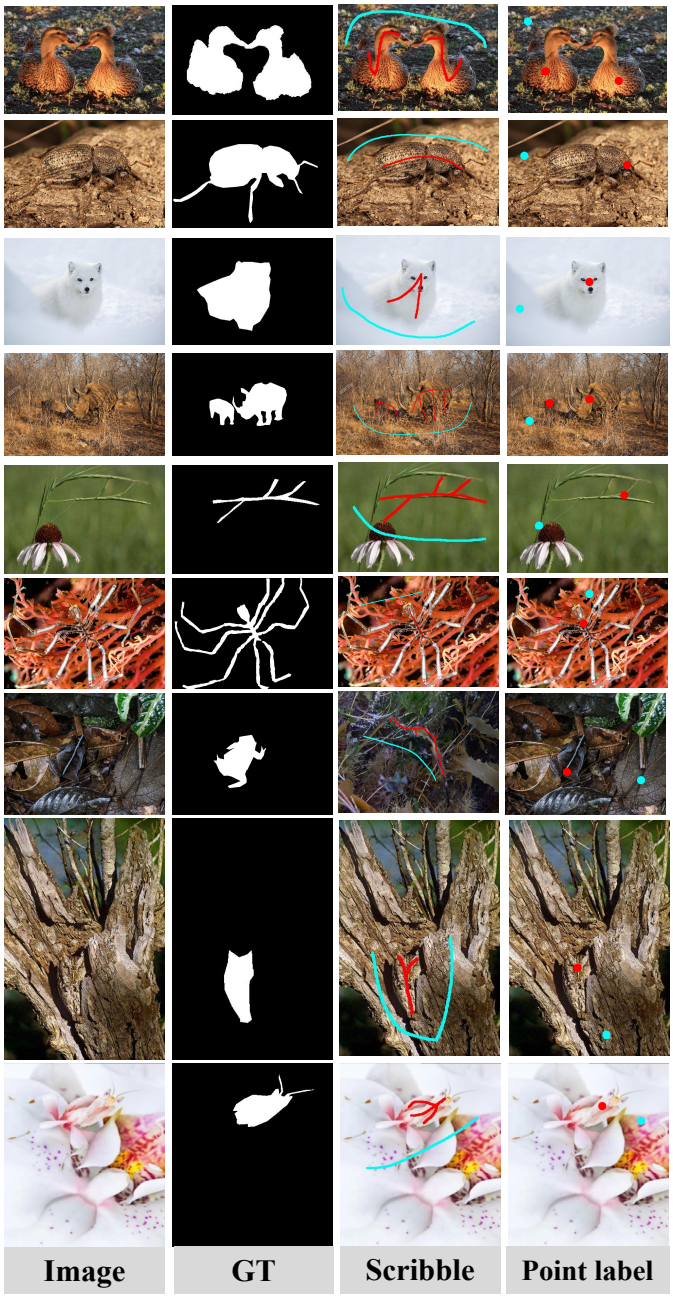


**Fig. 4:** Examples of PCOD dataset. It includes many categories of animals in challenging scenarios.

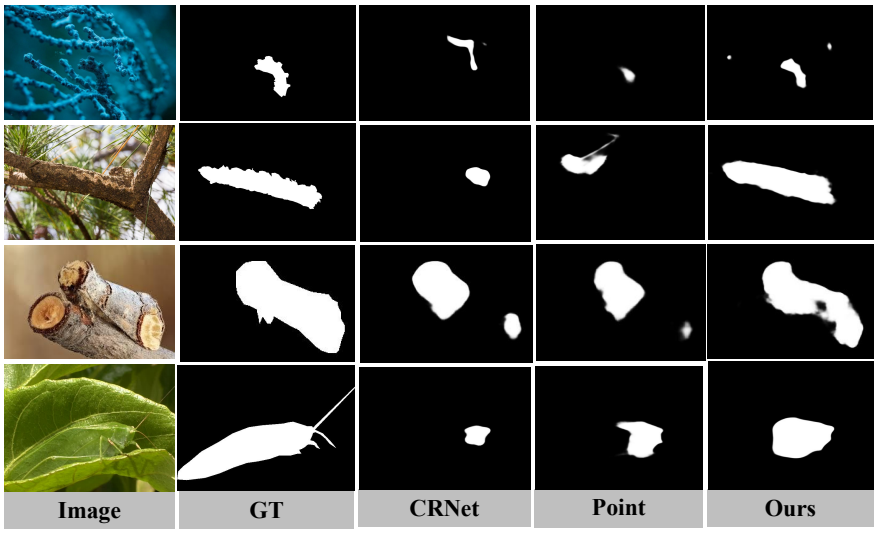


**Fig. 5:** The issue of WSCOD, part domination: focusing on discriminative parts of the object. Compared to directly using scribble or point supervision, our approach significantly alleviates the part domination issue.

only one point is annotated for each camouflaged object and background in an image. See Figure 4 for examples in our dataset (we exaggerate the size of the labeled position in visualization).

### 3.2 Visualization of Part Domination in COD.

As shown in Figure 5, It is observed that weakly-supervised COD still exhibits part domination (focusing on discriminative parts of the objects), despite not including category information. By proposing the attention regulator, our method greatly improves this issue.

### 3.3 Visualization in Hint Area Generator.

We visualize the result of main steps in hint area generator, including initial small squares, prediction maps, pre-processed circles, and final circles, as shown in Figure 6.

### 3.4 Replacing $\tau$ Hyperparameters with K-means.

As shown in Table 3, the $\tau$ can be replaced with unsupervised k-means clustering method. There is no impact on the performance of the model.

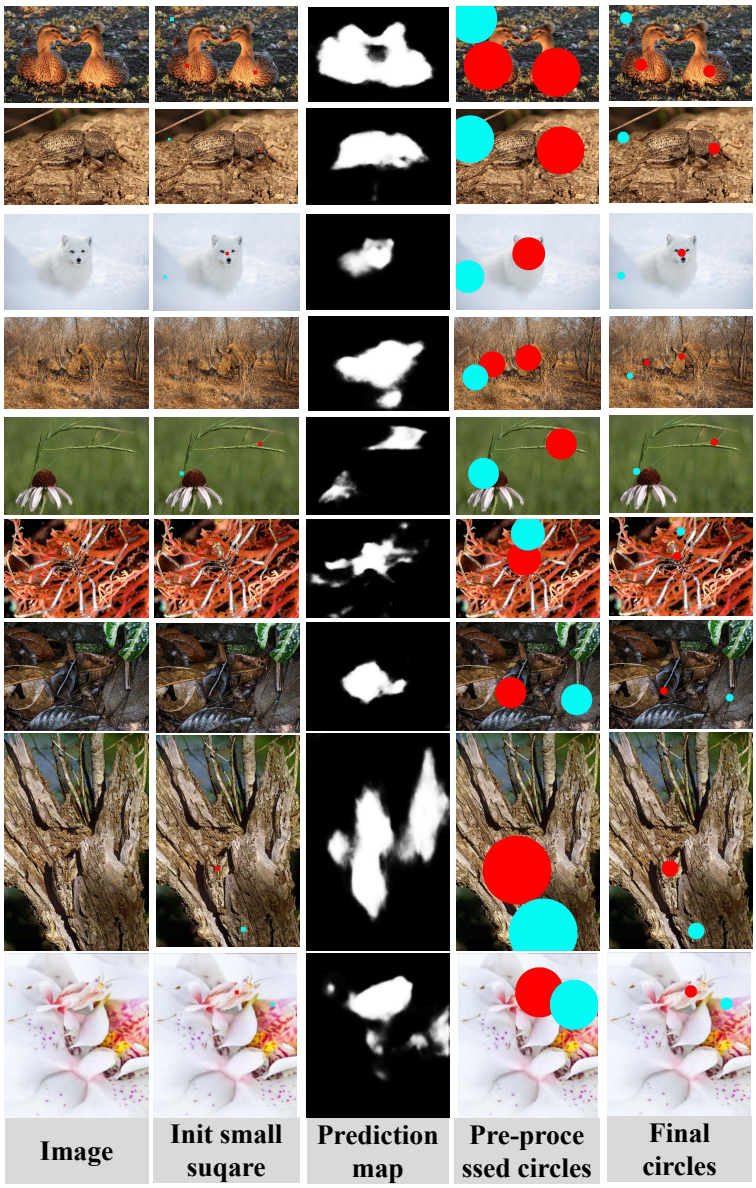


**Fig. 6:** Visualization in the hint area generator.

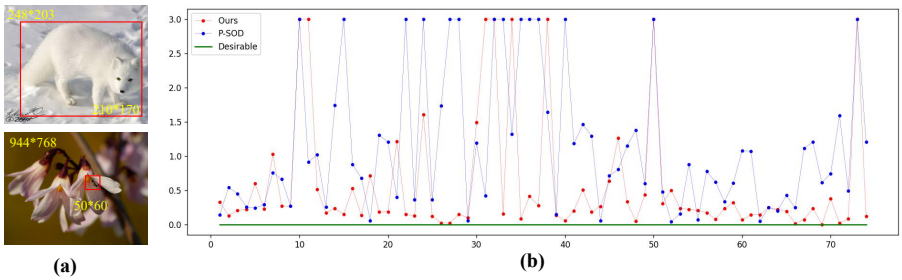


**Fig. 7:** (a) PSOD's incorrect prior: larger images result in smaller estimated objects. (b) A comparison between our prior and PSOD's prior.

**Table 1:** Quantitative comparison with state-of-the-arts on four benchmarks. "F", "U", "S", "P" denote Fully-supervised, Unsupervised, Scribble-supervised, and Point-supervised methods, respectively. The best are highlighted in **bold**.

| Methods | Sup. | CAMO (250) | | | | COD10K (2026) | | | | NC4K (4121) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE ↓ | $S_m$ ↑ | $E_m$ ↑ | $F_\beta^w$ ↑ | MAE ↓ | $S_m$ ↑ | $E_m$ ↑ | $F_\beta^w$ ↑ | MAE ↓ | $S_m$ ↑ | $E_m$ ↑ | $F_\beta^w$ ↑ |
| F3Net [23] | F | 0.109 | 0.711 | 0.741 | 0.564 | 0.051 | 0.739 | 0.795 | 0.544 | 0.069 | 0.782 | 0.825 | 0.706 |
| CSNet [6] | F | 0.092 | 0.771 | 0.795 | 0.641 | 0.047 | 0.778 | 0.809 | 0.569 | 0.061 | 0.819 | 0.845 | 0.748 |
| ITSD [33] | F | 0.102 | 0.750 | 0.779 | 0.610 | 0.051 | 0.767 | 0.808 | 0.557 | 0.064 | 0.811 | 0.845 | 0.729 |
| MINet [16] | F | 0.090 | 0.748 | 0.791 | 0.637 | 0.042 | 0.77 | 0.832 | 0.608 | - | - | - | - |
| PraNet [4] | F | 0.094 | 0.769 | 0.825 | 0.663 | 0.045 | 0.789 | 0.861 | 0.629 | - | - | - | - |
| UCNet [30] | F | 0.094 | 0.739 | 0.787 | 0.640 | 0.042 | 0.776 | 0.857 | 0.633 | 0.055 | 0.813 | 0.872 | 0.777 |
| SINet [3] | F | 0.092 | 0.745 | 0.804 | 0.644 | 0.043 | 0.776 | 0.864 | 0.631 | 0.058 | 0.808 | 0.871 | 0.723 |
| MGL-R [29] | F | 0.088 | 0.775 | 0.812 | 0.673 | 0.035 | 0.814 | 0.851 | 0.666 | 0.052 | 0.833 | 0.867 | 0.740 |
| PFNet [13] | F | 0.085 | 0.782 | 0.841 | 0.695 | 0.040 | 0.800 | 0.877 | 0.660 | 0.053 | 0.829 | 0.887 | 0.745 |
| UJSC [10] | F | 0.073 | 0.800 | 0.859 | 0.728 | 0.035 | 0.809 | 0.884 | 0.684 | 0.047 | 0.842 | 0.898 | 0.771 |
| UGTR [27] | F | 0.086 | 0.784 | 0.822 | 0.684 | 0.036 | 0.817 | 0.852 | 0.666 | 0.052 | 0.839 | 0.874 | 0.747 |
| ZoomNet [15] | F | 0.066 | 0.820 | 0.892 | 0.752 | 0.029 | 0.838 | 0.911 | 0.729 | 0.043 | 0.853 | 0.896 | 0.784 |
| DUSD [32] | U | 0.166 | 0.551 | 0.594 | 0.308 | 0.107 | 0.580 | 0.646 | 0.276 | - | - | - | - |
| USPS [14] | U | 0.207 | 0.568 | 0.641 | 0.399 | 0.196 | 0.519 | 0.536 | 0.265 | - | - | - | - |
| SAM [9] | U | 0.132 | 0.684 | 0.687 | 0.606 | 0.050 | 0.783 | 0.798 | **0.701** | 0.078 | 0.767 | 0.776 | 0.696 |
| SS [31] | S | 0.118 | 0.696 | 0.786 | 0.562 | 0.071 | 0.684 | 0.770 | 0.461 | - | - | - | - |
| SCSOD [28] | S | 0.102 | 0.713 | 0.795 | 0.618 | 0.055 | 0.710 | 0.805 | 0.546 | - | - | - | - |
| CRNet [8] | S | 0.092 | 0.735 | 0.815 | 0.641 | 0.049 | 0.733 | 0.832 | 0.576 | 0.063 | 0.775 | 0.855 | 0.688 |
| Ours | P | 0.074 | 0.798 | 0.872 | 0.727 | 0.042 | 0.784 | 0.859 | 0.650 | 0.051 | 0.822 | 0.889 | 0.748 |
| Ours + Scribble | S | **0.065** | **0.816** | **0.894** | **0.761** | **0.032** | **0.810** | **0.899** | **0.709** | **0.042** | **0.836** | **0.908** | **0.787** |

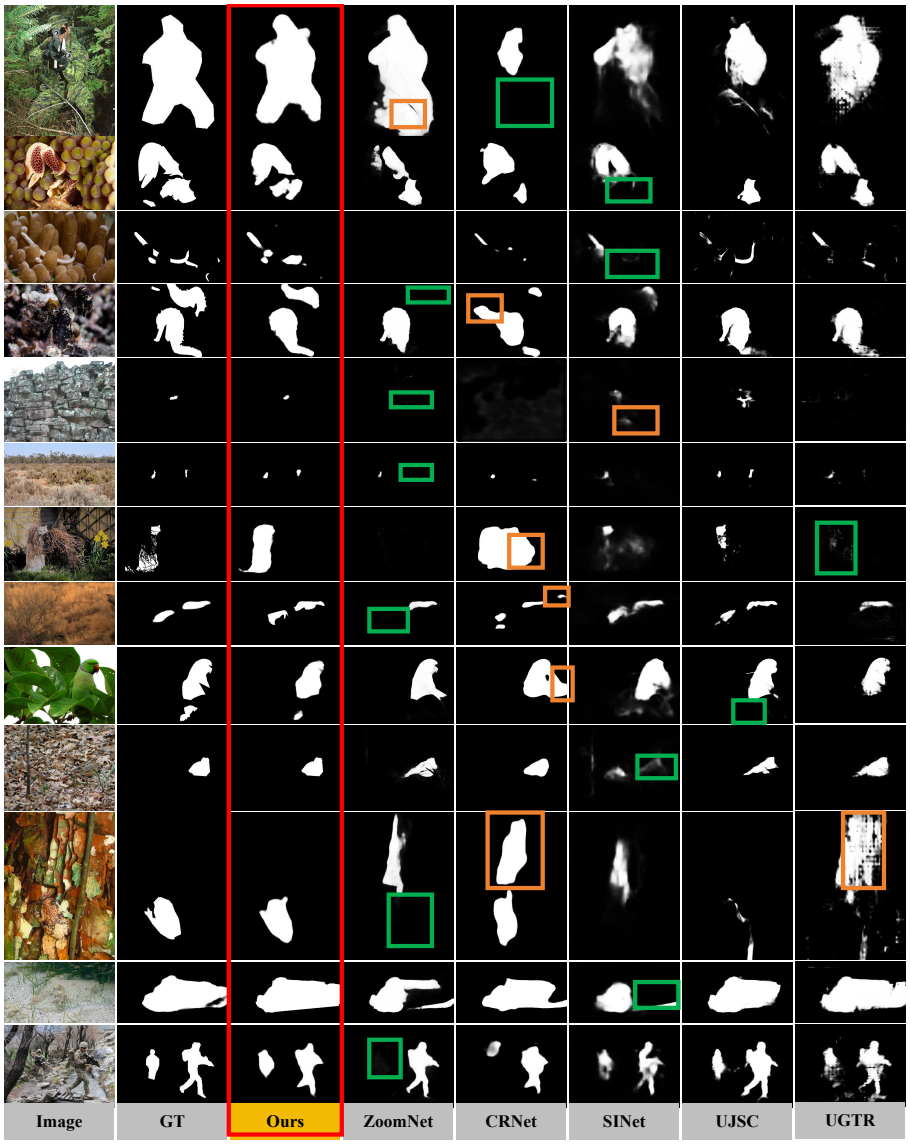


**Fig. 8:** Qualitative comparison of our method with state-of-the-arts methods. The red box represents the prediction results obtained by our method, the green box represents the missing parts in the prediction results, and the orange box represents the incorrectly predicted parts in results.

**Table 2:** Quantitative comparison with state-of-the-arts on five SOD benchmarks. "S", "P" denote scribble-supervised, and point-supervised methods, respectively. **Red** and **blue** represent the first and second best performing algorithms, respectively.

| Methods | Sup. | ECSSD | | | DUT-O | | | HKU-IS | | | DUTS-TE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE ↓ | $S_m$ ↑ | $F_\beta^{max}$ ↑ | MAE ↓ | $S_m$ ↑ | $F_\beta^{max}$ ↑ | MAE ↓ | $S_m$ ↑ | $F_\beta^{max}$ ↑ | MAE ↓ | $S_m$ ↑ | $F_\beta^{max}$ ↑ |
| RAS [1] | F | 0.056 | 0.893 | 0.921 | 0.062 | 0.814 | 0.786 | 0.045 | 0.887 | 0.913 | 0.059 | 0.839 | 0.831 |
| $R^3$Net [2] | F | 0.056 | 0.903 | 0.925 | 0.071 | 0.818 | 0.788 | 0.048 | 0.892 | 0.910 | 0.066 | 0.836 | 0.824 |
| DGR [20] | F | 0.041 | 0.903 | 0.922 | 0.062 | 0.806 | 0.774 | 0.036 | 0.892 | 0.910 | 0.050 | 0.842 | 0.828 |
| PiCANet [12] | F | 0.046 | 0.917 | 0.935 | 0.065 | 0.832 | 0.803 | 0.043 | 0.904 | 0.919 | 0.051 | 0.869 | 0.860 |
| MLMS [24] | F | 0.045 | 0.911 | 0.928 | 0.064 | 0.809 | 0.774 | 0.039 | 0.907 | 0.921 | 0.049 | 0.862 | 0.852 |
| AFNet [5] | F | 0.042 | 0.913 | 0.935 | 0.057 | 0.826 | 0.797 | 0.036 | 0.905 | 0.923 | 0.046 | 0.867 | 0.863 |
| BASNet [18] | F | 0.037 | 0.916 | 0.943 | 0.057 | 0.836 | 0.805 | 0.032 | 0.909 | 0.928 | 0.048 | 0.866 | 0.859 |
| MFNet [17] | S | 0.084 | 0.834 | 0.879 | 0.087 | 0.741 | 0.706 | 0.059 | 0.846 | 0.876 | 0.076 | 0.774 | 0.770 |
| WSSA [31] | S | 0.059 | 0.865 | 0.888 | 0.068 | 0.784 | 0.753 | 0.047 | 0.864 | 0.880 | 0.062 | 0.803 | 0.788 |
| SCSOD [28] | S | 0.049 | 0.881 | 0.914 | **0.060** | 0.811 | 0.782 | 0.038 | 0.882 | 0.908 | 0.049 | 0.841 | 0.844 |
| PSOD [7] | P | **0.036** | **0.913** | **0.935** | 0.064 | **0.824** | **0.808** | **0.033** | **0.901** | **0.923** | **0.045** | **0.853** | **0.858** |
| Ours | P | **0.045** | **0.895** | **0.915** | **0.059** | **0.826** | **0.824** | **0.031** | **0.906** | **0.926** | **0.042** | **0.864** | **0.861** |

**Table 3:** Ablation study of replacing $\tau$ hyperparameters with k-means.

| Type | MAE↓ | $S_m$ ↑ | $E_m$ ↑ | $F_\beta^w$ ↑ |
|---|---|---|---|---|
| K-means | 0.076 | 0.787 | 0.867 | 0.724 |
| $\tau$ | 0.076 | 0.790 | 0.866 | 0.724 |

### 3.5 Detailed Quantitative Comparison.

**Results on Scribble Datasets.** We train our model using scribble labels and still achieve excellent results, as shown in Table 1.
**Results on SOD Datasets.** We compare the proposed model with existing methods in salient object detection. All the results are listed in Table 2. Our model show outstanding performance in SOD task, which verifies that the proposed model can deal with the more general binary segmentation task.

### 3.6 Detailed Qualitative Comparison.

To further demonstrate the superior performance of our approach, we showcase additional results, as shown in Figure 8. It is clearly observed that our method can recover more complete regions of objects.

### 3.7 The Reliability of Estimation.

As shown in Figure 7 (a), PSOD [7] uses the incorrect prior that the image pixel size is approximately equal to object pixel size. For the inaccurate prior, we use the count of the prediction map pixel to approximate the object pixel size as a prior instead of the prior in PSOD. We conducted tests by randomly

selecting 5% of the images from the training set and used both PSOD and our estimation method for object size estimation. When the estimated object size (in pixels) matches the actual object size, the ratio between the two is the same, and sample points lie close to the green line (ideal line). As shown in Figure 7 (b), experimental results demonstrate that our estimation is more accurate than PSOD.

## 4 Future.

Our model, a weakly supervised model, performs on par with most fully supervised models in handling challenging COD tasks with only single-point annotations. It can even surpass state-of-the-art fully supervised models when utilizing scribble annotations. We think that there is much room for developing models to bridge the gap toward better data-efficient learning, such as self- and semi-supervised learning.

## References

1. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 234–250 (2018)
2. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th international joint conference on artificial intelligence. pp. 684–690. AAAI Press Menlo Park, CA, USA (2018)
3. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2777–2787 (2020)
4. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
5. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1623–1632 (2019)
6. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: European Conference on Computer Vision. pp. 702–721. Springer (2020)
7. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-supervised salient object detection using point supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 670–678 (2022)
8. He, R., Dong, Q., Lin, J., Lau, R.W.: Weakly-supervised camouflaged object detection with scribble annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 781–789 (2023)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

10. Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., Dai, Y.: Uncertainty-aware joint salient object and camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10071–10081 (2021)
11. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5455–5463 (2015)
12. Liu, N., Han, J., Yang, M.H.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3089–3098 (2018)
13. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8772–8781 (2021)
14. Nguyen, T., Dax, M., Mummadi, C.K., Ngo, N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction via self-supervision. Advances in Neural Information Processing Systems **32** (2019)
15. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 2160–2170 (2022)
16. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9413–9422 (2020)
17. Piao, Y., Wang, J., Zhang, M., Lu, H.: Mfnet: Multi-filter directive network for weakly supervised salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4136–4145 (2021)
18. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7479–7489 (2019)
19. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017)
20. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3127–3135 (2018)
21. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
22. Wei, J., Wang, Q., Li, Z., Wang, S., Zhou, S.K., Cui, S.: Shallow feature matters for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5993–6001 (2021)
23. Wei, J., Wang, S., Huang, Q.: $F^3$net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. pp. 12321–12328 (2020)
24. Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., Ding, E.: A mutual learning method for salient object detection with intertwined multi-supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8150–8159 (2019)
25. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1155–1162 (2013)

26. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3166–3173 (2013)
27. Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., Fan, D.P.: Uncertainty-guided transformer reasoning for camouflaged object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4146–4155 (2021)
28. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: Proceedings of the AAAI conference on artificial intelligence. pp. 3234–3242 (2021)
29. Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., Fan, D.P.: Mutual graph learning for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12997–13007 (2021)
30. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8582–8591 (2020)
31. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12546–12555 (2020)
32. Zhang, J., Zhang, T., Dai, Y., Harandi, M., Hartley, R.: Deep unsupervised saliency detection: A multiple noisy labeling perspective. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9029–9038 (2018)
33. Zhou, H., Xie, X., Lai, J.H., Chen, Z., Yang, L.: Interactive two-stream decoder for accurate and fast saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9141–9150 (2020)