Uncertainty-aware Joint Salient Object and Camouflaged Object Detection–Supplementary Material

Aixuan Li[1,‡] Jing Zhang[2,3,‡] Yunqiu Lv[1] Bowen Liu[1] Tong Zhang[4] Yuchao Dai[1]✉
[1] Northwestern Polytechnical University, China [2] Australian National University, Australia
[3] CSIRO, Australia [4] EPFL, Switzerland
‡ Equal contributions; ✉ Corresponding author:daiyuchao@nwpu.edu.cn

Abstract

In this supplementary material, we provide our network structure in detail, more samples we selected from the camouflage training dataset to the saliency training dataset, and predictions of our network for the connection modeling dataset, e.g. *PASCAL VOC 2007 dataset [1].*

1. Network Structure

In our joint salient object detection and camouflaged object detection pipeline, we introduced an uncertainty-aware framework, which consists of a "Feature encoder", a "Prediction decoder", a "Similarity measure module", and a "Confidence estimation module".

The "Feature encoder" is based on ResNet50[4], which includes a saliency encoder and a camouflage encoder. The saliency encoder and the camouflage encoder have the same network structure, and they output four groups of features: $F_{\alpha_s} = \{f_1^s, f_2^s, f_3^s, f_4^s\}$ and $F_{\alpha_c} = \{f_1^c, f_2^c, f_3^c, f_4^c\}$ respectively.

The input of the "Prediction decoder" is the output of the "Feature encoder", where COD and SOD share the same decoder G_β, and β is the parameter set of the prediction decoder. The structure of the "Prediction decoder" is shown in Fig. 1.

The "Similarity measure module" is used to model the connection between SOD and COD. We utilize the PASCAL VOC 2007 dataset [1] as a connection modeling dataset. The input of "Similarity measure module" is the encoded features of the PASCAL VOC 2007 dataset from both SOD and COD encoders: $F^p_{\alpha_s} = \{f^p_{s1}, f^p_{s2}, f^p_{s3}, f^p_{s4}\}$ and $F^p_{\alpha_c} = \{f^p_{c1}, f^p_{c2}, f^p_{c3}, f^p_{c4}\}$ respectively. It produces the latent feature: $f^p = S_\theta(F^p_{\alpha_s}, F^p_{\alpha_c})$, θ is parameter set of the "Similarity measure module". We generate their latent space feature separately through similarity measure and calculate their cosine similarity. The structure of "Similarity measure module" is shown in Fig. 2.

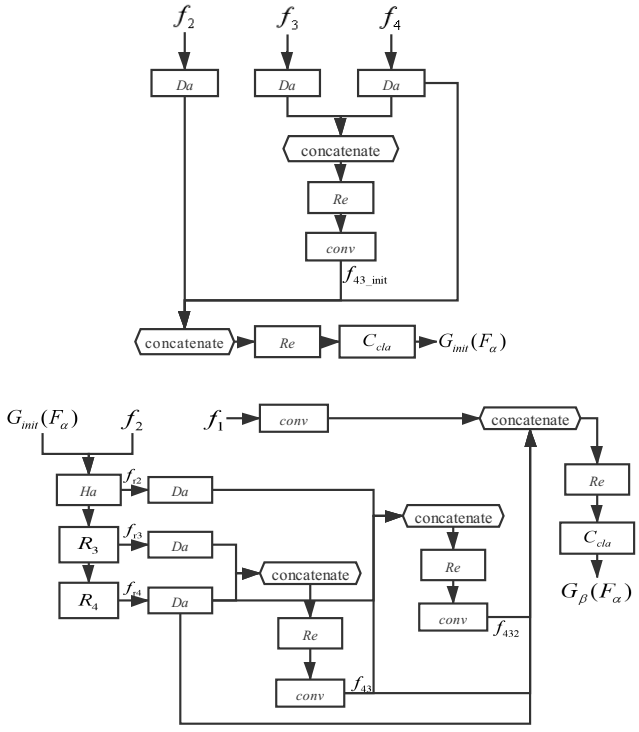

Figure 1. The network of the "Prediction decoder". Da is the dual attention module [3], Ha is the holistic attention module[6], Re is the residual channel attention module[8]. $conv$ is the 3×3 convolutional layer of output channel size $C = 32$, and C_{cla} map the feature map to one channel prediction. R_3 and R_4 are the ResNet50 backbone convolutional layers of channel size 1024 and 2048 respectively.

The "Confidence estimation module" is proposed to explicitly model the confidence of network predictions, which takes model predictions ($G_\beta(F_{\alpha_s})$, $G_\beta(F_{\alpha_c})$) (or ground truth maps) as input, and produce pixel-wise confidence map ($D^f_\gamma(G_\beta(F_{\alpha_s}))$, $D^f_\gamma(G_\beta(F_{\alpha_c}))$), where γ is network parameter set of the confidence estimation module. The network structure is shown in Fig. 3.

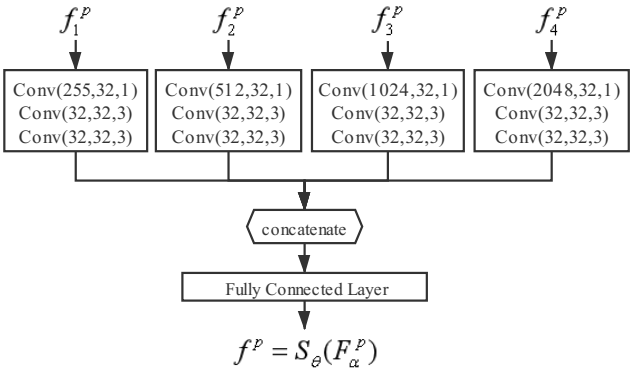

Figure 2. The network of the "Similarity measure module". $f_k^p, k = 1, ..., 4$ are the feature from SOD encoder and COD encoder, the fully connected layer is to obtain the latent space, we set the dimension of the latent space as $K = 700$.

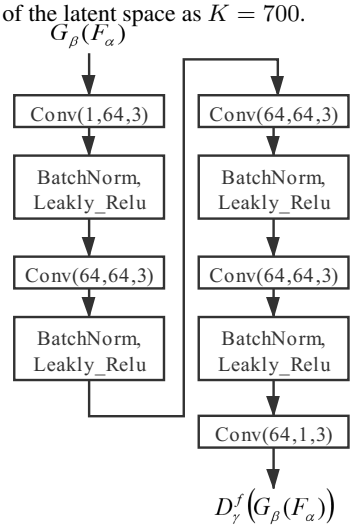

Figure 3. The network of the "Confidence estimation module".

2. Data interaction

We find that the SOD task and COD task include some overlap images, which are both salient and camouflaged. We would like to argue that these easy camouflaged images could be regarded as hard positive samples for the SOD task. We then choose 400 images from the COD training dataset [2], which achieves the smallest MAE by testing it using a trained SOD model [7]. We replace them with randomly selected samples from the SOD training dataset [5]. It is proved that these hard samples could provide robustness for SOD model. We show some selected COD samples in Fig. 4. The first and third lines are the selected images from the COD training dataset[2], and the second and fourth lines are the corresponding ground-truth.

Note that, although the easy samples from the COD training dataset can be treated as hard samples for the SOD training dataset, the opposite way cannot work. The main reason is that most salient objects can appear in our surroundings, while the habitats of the camouflaged objects usually are far away from us, which makes the hard salient samples not appropriate to serve as easy camouflage samples. Meanwhile, the hard saliency samples are those with a complex background, while this attribute does not always indicate the existence of the camouflaged objects.

3. Connection modeling dataset

In the "Similarity measure module", we introduced the PASCAL VOC 2007 dataset [1] to model the connection between SOD and COD. We argue that it could lead to the latent features different from each task and force the two tasks to focus on different regions of the image. We further show the detected regions of PASCAL VOC 2007 dataset from SOD and COD tasks in Fig. 5

References

[1] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88:303–338, 06 2010. 1, 2, 3

[2] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2777–2787, 2020. 2

[3] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[5] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 136–145, 2017. 2

[6] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1

[7] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019. 2

[8] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1

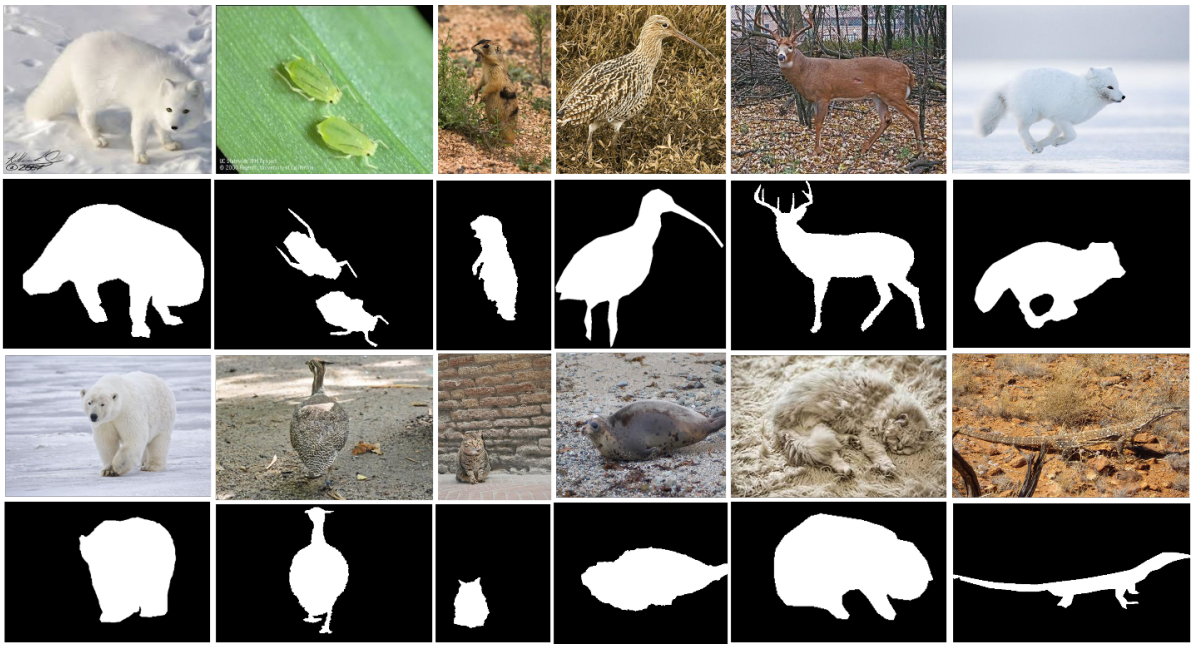

Figure 4. Selected easy samples from the COD training dataset. We argue that these samples could be used as the hard samples for SOD task, and make the SOD model more robust.

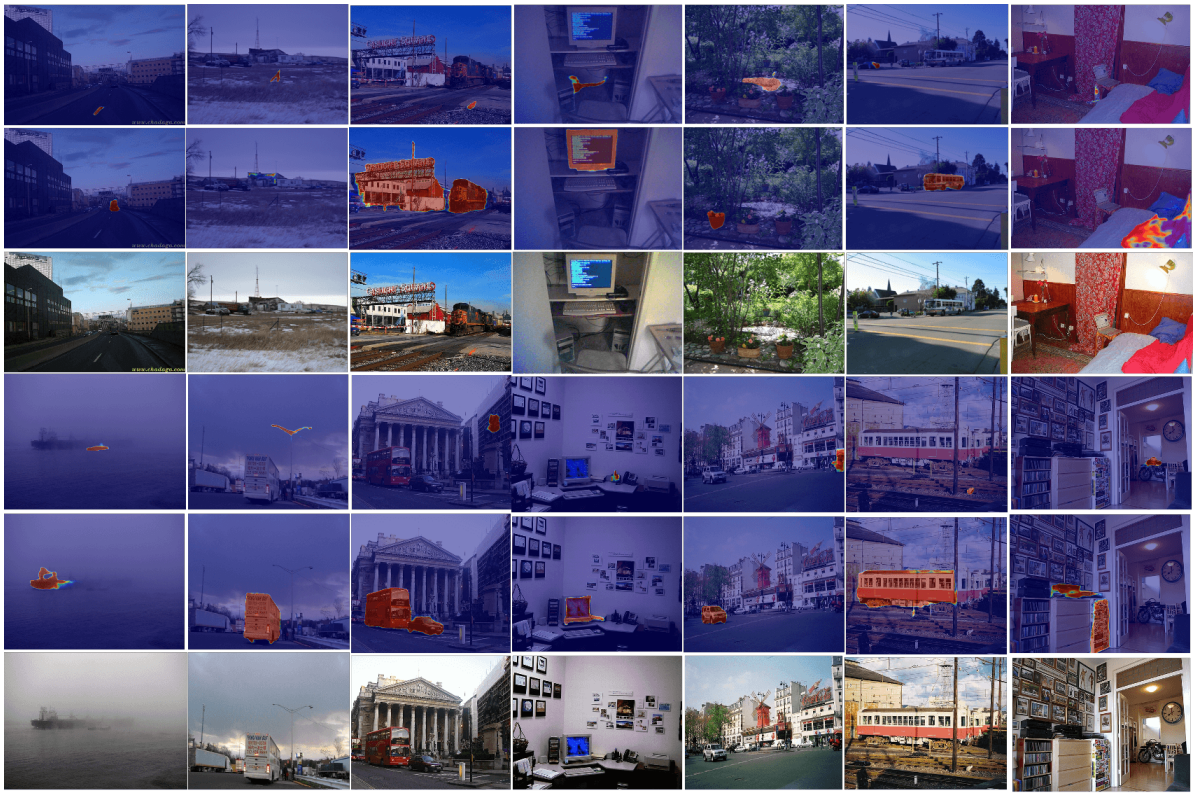

Figure 5. The detected region from the camouflage encoder branch (first, fourth rows), the saliency encoder branch (second, fifth rows) and the original RGB image from PASCAL VOC 2007 dataset [1] (third, sixth rows).