Weakly-Supervised Salient Instance Detection (Supplementary Materials)

Xin Tian[1][2]
xtian@mail.dlut.edu.cn
Ke Xu[1][2]
kkangwing@mail.dlut.edu.cn
Xin Yang[1]
xinyang@dlut.edu.cn
Baocai Yin[1][3]
ybc@dlut.edu.cn
Rynson W.H. Lau[2]
rynson.lau@cityu.edu.hk

[1] Computer Science Department
Dalian University of Technology
Dalian, China

[2] Computer Science Department
City University of Hongkong
Hongkong SAR, China

[3] Pengcheng Lab
Shenzhen, China

1 Overview

In this Supplemental, we provides more results that is organised as below:

- Section 2: more internal analysis relating to our Double Attention module and the backbone used in proposed model;
- Section 3: more qualitative results of the predicted boundaries, centroids, and salient regions as shown in Figures 1 and 2;
- Section 4: more qualitative results of salient instance detection, compared with the modified weakly-supervised baselines (*i.e.*, PRM+D [2], DeepMask [8], C2SNet [6], NLDF [7], and IRN [1]) and existing fully-supervised methods (*i.e.*, S4Net[3], and MAP [10]) as shown in Figures 3, 4, and 5. As of today, the codes for MSRNet [5] are still not available. Therefore, we do not visually compare to it, same as the main paper.

2 Internal analysis

2.1 Double Attention (DA) module

We investigate the design choices of our Double Attention module against its several variants, as reported in Table 1. First, we study the case of using one kind of attention mechanisms only in row 1 and 2. We can see that the performance drops about 2 percent compared to our DA, this is due to independent attention mechanism provides relative weak context information. Second, we inquire the comparison between cascade and parallel attention mechanisms. Row 3, 4 and 5 show that our parallel one outperforms the cascade attention designs with about 1% higher numbers. The reason is that cascade connection may lose the learned context information of the former attention mechanism.

method	mAP@0.5↑	mAP@0.7↑
using channel-wise attention only	59.8%	45.0%
using spatial-wise attention only	60.3%	45.4%
using cascade attention (s→c)	60.9%	46.2%
using cascade attention (c→s)	61.1%	46.5%
Our DA (parallel attention)	**61.9%**	**47.2%**

Table 1: Evaluation of different designs of the DA module. s→c represents using spatial-wise attention before channel-wise attention, while c→s represents the reverse connection. The best performance among different designs is marked in **bold**.

2.2 different backbones

We also study the performance influence of using different backbones for our model, as shown in Table 2. We select three popular backbones, *i.e*. VGG19 [9], ResNet50 [4], and ResNet101 [4], for the comparison. VGG19 has relative shallow feature layers and no residual connection, making it not as efficient as ResNet for our task, as shown in the row 1 (about 1% lower). Even the ResNet101 has deeper feature layers than ResNet50, it might be redundant for our task, as shown in the row 2 (0.5-0.7% lower). As a result, we choose ResNet50 as the backbone for our model.

backbone	mAP@0.5↑	mAP@0.7↑
VGG19 [9]	60.8%	45.7%
ResNet101 [4]	61.4%	46.5%
ResNet50 [4] (finally used in WSID-Net)	**61.9%**	**47.2%**

Table 2: Evaluation of the SID performance of using different backbones for our WSID-Net. The best performance among different choices is marked in **bold.**

3 Output quality

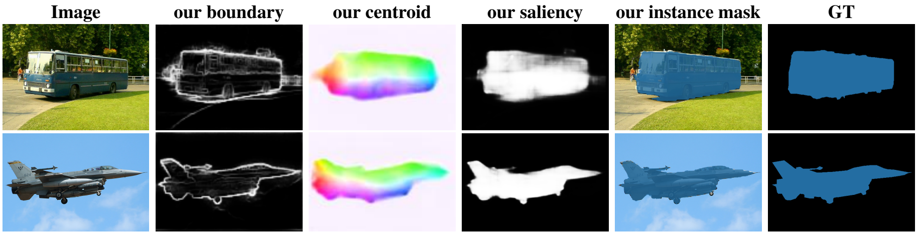

Figure 1: Qualitative results of predicted boundaries, centroids, salient regions, and inferred instance masks of our WSID-Net.

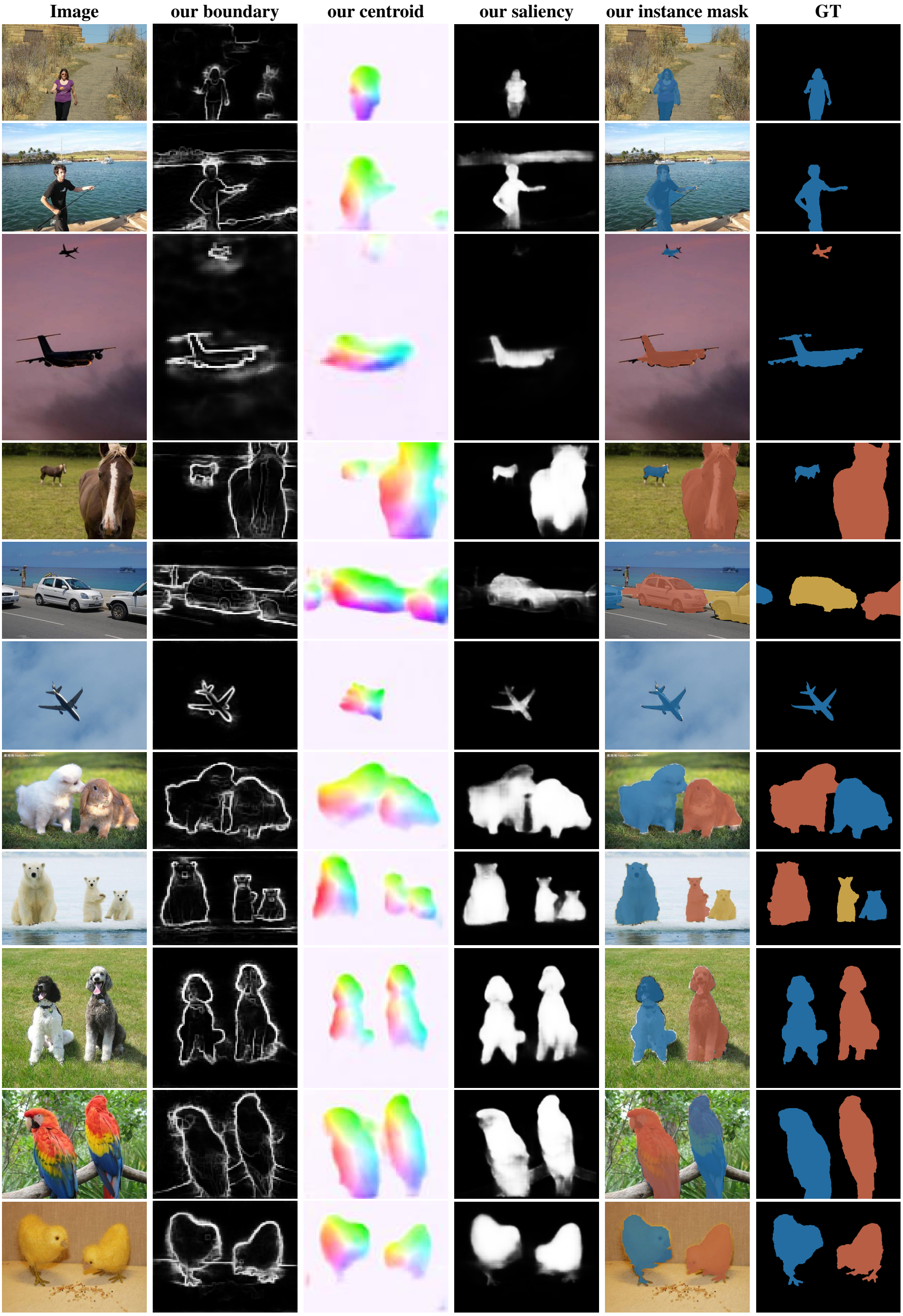

Figure 2: Qualitative results of predicted boundaries, centroids, salient regions, and inferred instance masks of our WSID-Net.

4 Qualitative comparison

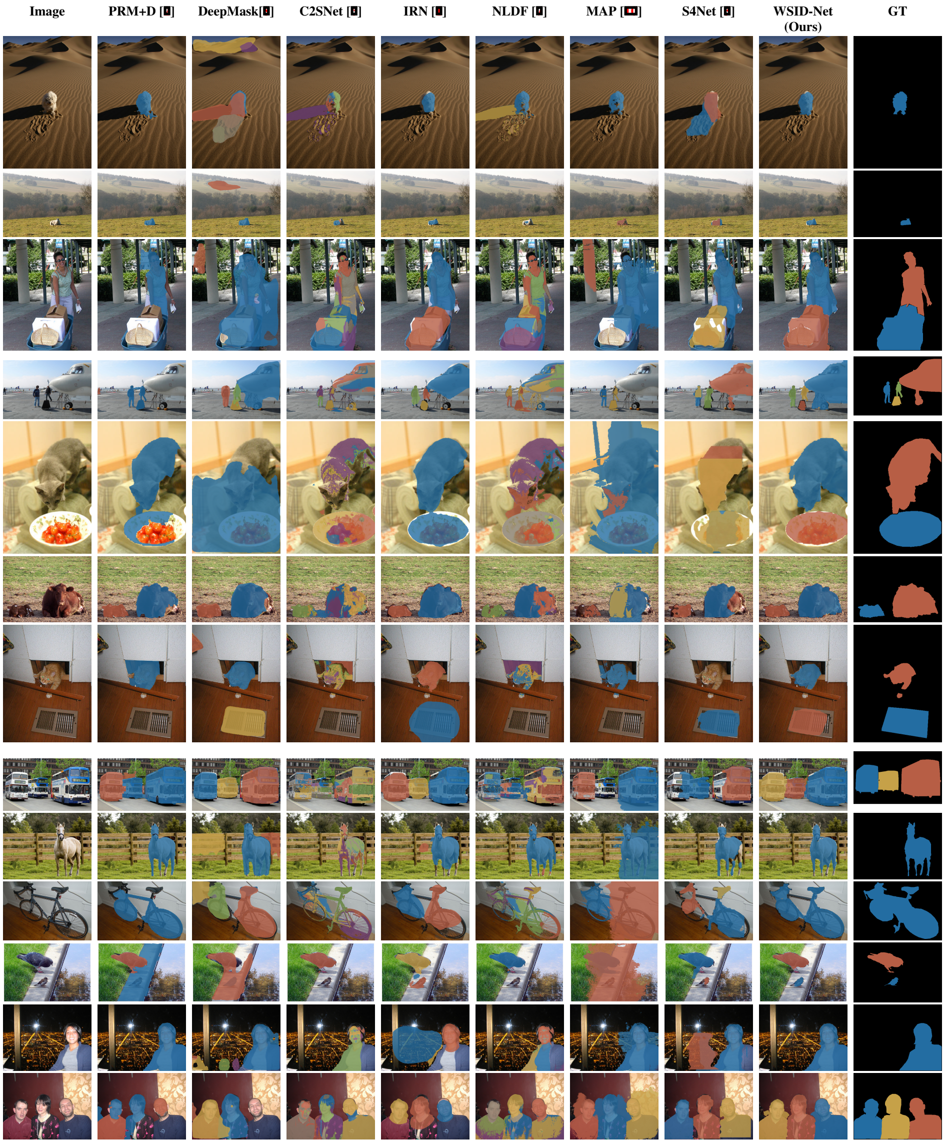

Figure 3: Qualitative results of our method, compared with existing fully-supervised methods (S4Net[3] and MAP [10]) and modified baselines (PRM+D [2], DeepMask [8], C2SNet [6], NLDF [7], and IRN [1]). Refer to Section 4.3 and Table 1 in main paper on how we modify and train these baselines, in order to perform appropriate comparison.

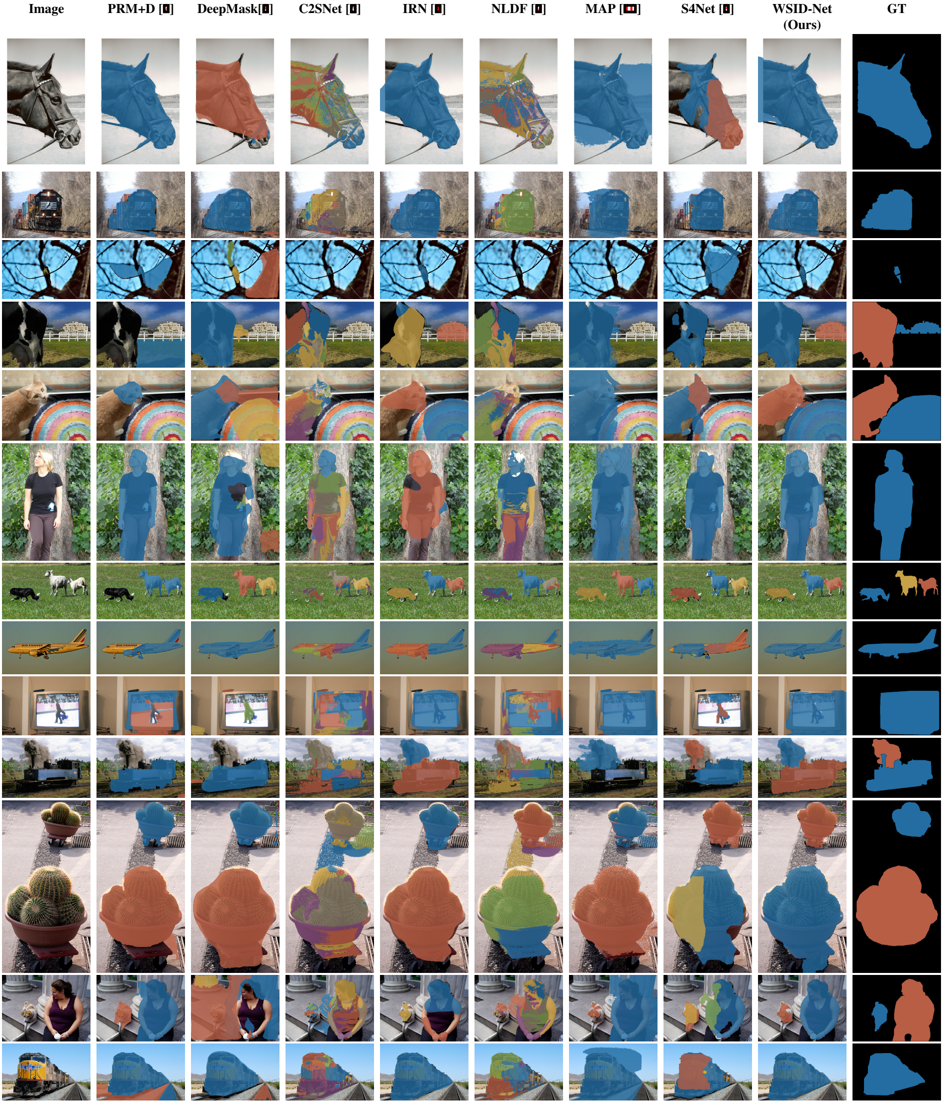

Figure 4: Qualitative results of our method, compared with existing fully-supervised methods (S4Net[3] and MAP [10]) and modified baselines (PRM+D [2], DeepMask [8], C2SNet [6], NLDF [7], and IRN [1]). Refer to Section 4.3 and Table 1 in main paper on how we modify and train these baselines, in order to perform appropriate comparison.

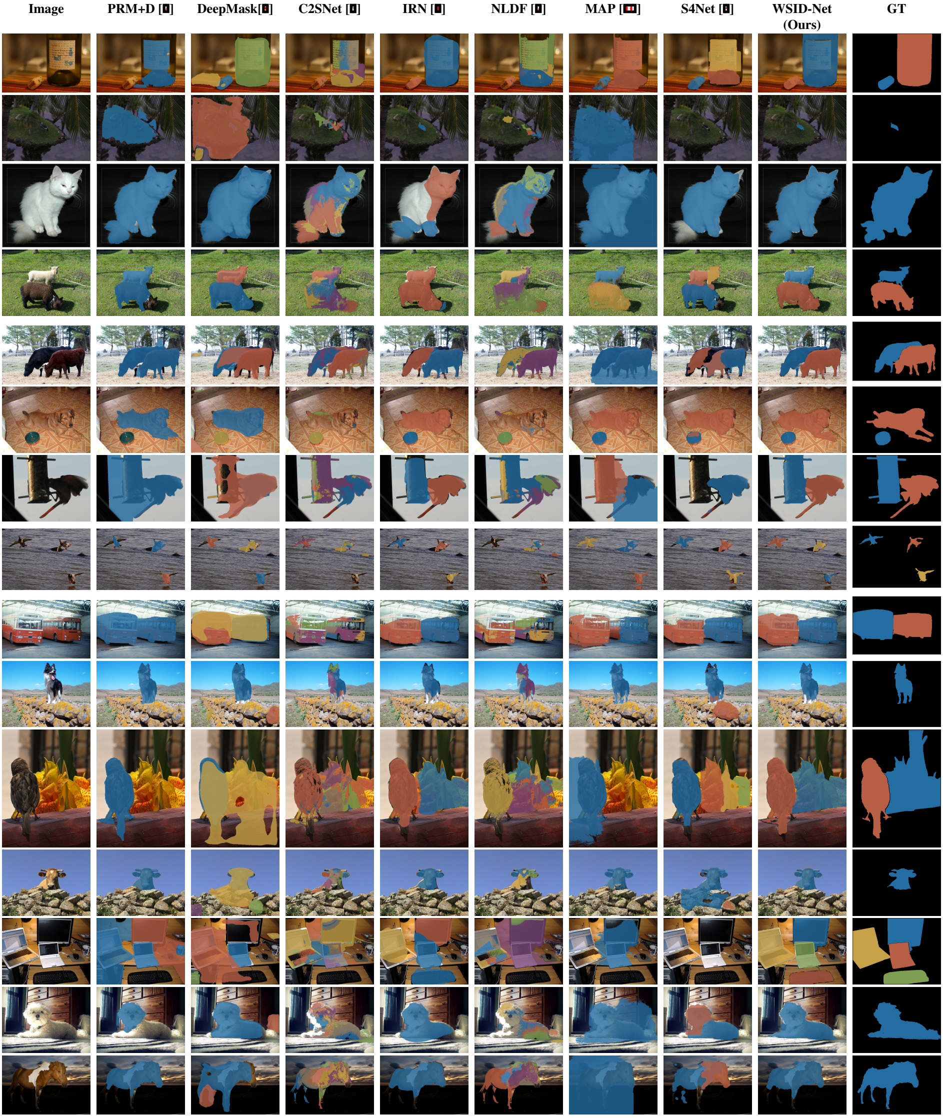

Figure 5: Qualitative results of our method, compared with existing fully-supervised methods (S4Net[3] and MAP [10]) and modified baselines (PRM+D [2], DeepMask [8], C2SNet [6], NLDF [7], and IRN [1]). Refer to Section 4.3 and Table 1 in main paper on how we modify and train these baselines, in order to perform appropriate comparison.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

[2] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *CVPR*, 2019.

[3] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[5] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017.

[6] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, 2018.

[7] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017.

[8] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NeurIPS*, 2015.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[10] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *CVPR*, 2016.