Supplementary Results:Object Counting and Instance Segmentation with Image-level Supervision

Hisham Cholakkal[1*] Guolei Sun[1*] Fahad Shahbaz Khan[1,2]
Ling Shao[1]
[1]Inception Institute of Artificial Intelligence, UAE
[2]Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Sweden
{hisham.cholakkal, guolei.sun, fahad.khan, ling.shao}@inceptioniai.org

1. Additional Ablation Studies

1.1. Using object count annotations beyond subitizing range:

Fig. 1 shows that the improvement obtained using exact count annotations for the entire range (IC, $\tilde{t} = \infty$) is marginal compared to the proposed image-level lower-count (ILC) supervision ($\tilde{t} = 5$). This is likely due to the joint optimization of the spatial and MSE loss terms within the subitizing range, which enables the density branch to predict accurate counts even beyond the subitizing range. In addition, the ranking loss helps to improve the model performance by penalizing under-counting beyond the subitizing range. Our approach provides an optimum balance between annotation cost and counting performance and hence can be easily extended to new datasets. Fig. 1 also shows that a single-branch architecture using an MSE loss for all counts, corresponding to glancing [2], leads to inferior results (see Tab. 1 in the main paper).

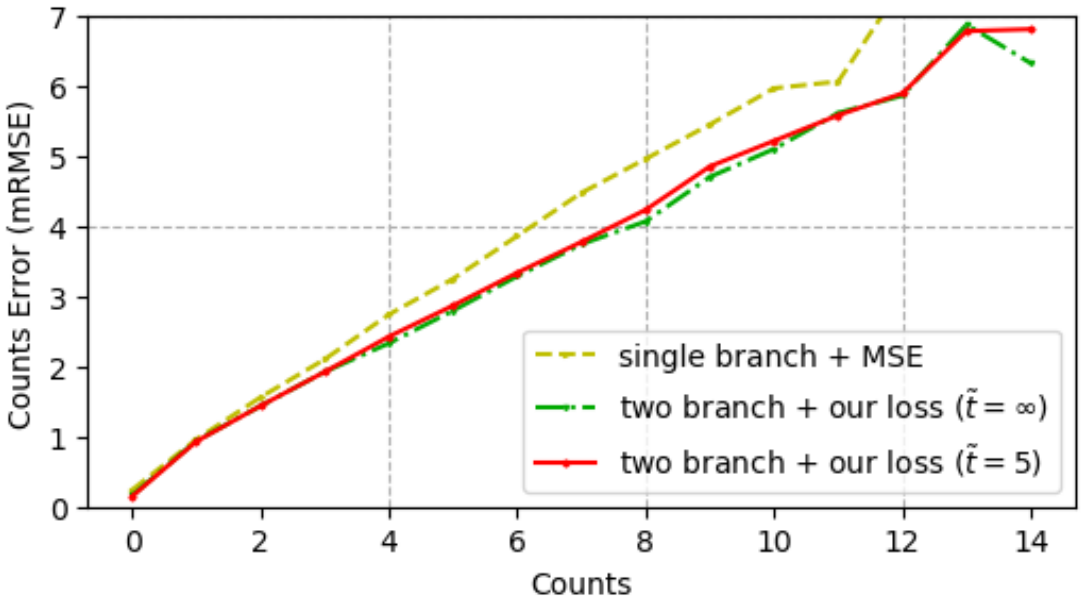

Figure 1: Counting performance comparison (in mRMSE) at different ground-truth count values on COCO. When training with reduced supervision (ILC, $\tilde{t}$=5), our two branch architecture with the proposed loss function achieves performance that is comparable to training with exact object count (IC, $\tilde{t} = \infty$).

*Equal contribution

Method	PRM [5]	PRM [5]+CSRnet [4]	Ours
mAP	26.8	27.3	**30.2**

Table 1: Our approach vs. a recent density estimation method, CSRnet [4], for improving SOTA instance segmentation (PRM).

1.2. Image-level supervised Instance segmentation with other density estimation methods

To the best of our knowledge, we are the first to introduce the idea of using density maps to improve image-level supervised instance segmentation in natural scenes. To adhere with the problem settings, an *image-level* supervised density map estimation method is desired. But, existing density map estimation methods [1,4], mostly target crowded surveillance scenes with one or few object categories and are typically trained using point-level supervision.

We perform an experiment by adapting a recent point-level supervised crowd density estimation approach, CSRnet [4], to produce all 20 PASCAL VOC category density maps for instance segmentation (see Tab. 1). The inferior results of CSRnet ($mAP^r_{0.5}$=27.3 vs. 30.2 of ours) align with our density map evaluation in Sec. 4.1 of main paper. CSRnet and other crowd density estimation methods [1,3] generally assume minimal intra-class variations in the object size or shape, and ground-truth masks are generated accordingly. This assumption often fails in natural scenes, causing inferior performance.

2. Failure cases

2.1. Common object counting

Figure 2: Failure cases of the proposed common object counting method on COCO count-test set. The ground-truth and our predictions are shown in black and green respectively. Column 2 shows under-counting of bench category while other three images shows over-counting for different object categories.

2.2. Instance Segmentation

Figure 3: Failure cases of the proposed instance segmentation method on challenging images from PASCAL VOC 2012 dataset. Top row shows the input image and the bottom row shows corresponding instance masks predicted by the proposed method. In the first image, the bird mask is extended to the background due to the poor contrast. Transparent bus regions in the second image and truncated cow instance in the fourth image causes false negatives in the instance predictions.

References

[1] X. Cao, Z. Wang, Y. Zhao, and F. Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018. 2

[2] P. Chattopadhyay, R. Vedantam, R. R. Selvaraju, D. Batra, and D. Parikh. Counting everyday objects in everyday scenes. In *CVPR*, 2017. 1

[3] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*. 2010. 2

[4] Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, 2018. 2

[5] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. In *CVPR*, 2018. 2