Supplementary Materials for The Devil is in the Points: Weakly Semi-Supervised Instance Segmentation via Point-Guided Mask Representation

Beomyoung Kim[1,2] Joonhyun Jeong[1,2] Dongyoon Han[3] Sung Ju Hwang[2]

NAVER Cloud, ImageVision[1] KAIST[2] NAVER AI Lab[3]

Appendix: Additional Experimental Details

Labeling Budget Calculation. Seminar works [2, 3] offered the annotation time of various labeling sources (*e.g.*, full mask, bounding box, point, image-level labels) on Pascal VOC dataset [7]. Since the COCO dataset [14] we used has more categories and instances per image than the VOC dataset, we estimate the labeling budget for the COCO dataset following their budget calculation method. The COCO 2017 trainset has a total of 80 categories and contains 118,287 images and 860,001 instances. Also, it has an average of 7.2 instances and 2.9 categories per image. By considering this statistic of COCO dataset, we calculate the labeling budget as follows:

- **Full mask**: $77.1 classes/img \times 1s/class + 7.2 inst/img \times 79s/mask =$ **645.9s/img**.
- **Bounding box**: $77.1 classes/img \times 1s/class + 7.2 inst/img \times 7s/bbox =$ **127.5s/img**.
- **Point**: $77.1 classes/img \times 1s/class + 2.9 classes/img \times 2.4s/point + (7.2 inst/img - 2.9 classes/img) \times 0.9s/point =$ **87.9s/img**.
- **Image-level**: $80 classes/img \times 1s/class =$ **80s/img**.

Input of MaskRefineNet. In this section, we further provide the details about the input sources for MaskRefineNet. After training the teacher network using the fully labeled data, we generate instance mask outputs for the point-guided filtered proposals (*i.e.*, true-positive proposals) using the trained teacher network. We treat the mask outputs as rough masks to be used as the input source of the MaskRefineNet. For each rough mask, we loosely crop each instance region in the input image, rough mask, and point heatmap. Specifically, after obtaining the bounding box from the rough mask using the min-max operations, we re-scale the size of the box to double, and then we use this box region as the cropping region. In addition, for the point heatmap, we encode each point to a 2-dimensional gaussian kernel with a sigma of 6, as done in [19, 22].

Input Size	AP	AP_{50}	AP_{75}
128×128	34.1	53.4	36.1
256×256	35.5	56.0	37.8
384×384	35.5	55.9	37.7

Table 1. **Effect of the input size of MaskRefineNet**. The APs are measured on COCO 2017 validation set.

Iterative	1%	2%	5%	10%	30%	50%	100%
	23.9	25.1	33.4	35.5	37.4	38.3	39.0
✓	25.6	26.0	34.5	35.9	37.6	38.3	39.0

Table 2. **Effect of iterative training strategy.** The APs are measured on COCO 2017 validation set according to COCO subsets.

We concatenate the three input sources (*i.e.*, cropped input image $\mathcal{R}^{H\times W\times 3}$, cropped rough mask $\mathcal{R}^{H\times W\times 1}$, and cropped point heatmap $\mathcal{R}^{H\times W\times C}$) to be the input tensor $\mathcal{R}^{H\times W\times(3+1+C)}$ of the MaskRefineNet, where C is the number of classes.

Appendix: Additional Analysis

Effect of the input size of MaskRefineNet. We originally set the input size of MaskRefineNet to 256×256. Here, we change the input size to verify its effect on the WSSIS result in table 1. For this, we train the MaskRefineNet using the input size of 128×128 or 384×384. We measure the AP result of the student network trained with the pseudo and full labels on the COCO 2017 validation set. Consequently, the 256×256 size yields the best AP score of 35.5% but its performance gap with the 384×384 size is marginal (35.5% vs 35.4%).

Effect of iterative training strategy. Some weakly-supervised methods [1, 11, 17] utilize iterative training strategy; after training the target network, they generate pseudo labels using the target network, and then they newly train the target network using the pseudo labels. This strat-

Method	Label Types	Budget (days) ↓	AP (%) ↑
Weakly-supervised Models			
BBTP [10]	$\mathcal{B}$ 100%	174.5	21.1
BBAM [12]	$\mathcal{B}$ 100%	174.5	25.7
BoxInst [16]	$\mathcal{B}$ 100%	174.5	33.2
BoxLevelSet [13]	$\mathcal{B}$ 100%	174.5	33.4
BoxTeacher [6]	$\mathcal{B}$ 100%	174.5	35.4
Point-sup [5]	$\mathcal{P}_{10}$ 100%	263.2	37.7
Weakly Semi-supervised Models			
Ours	$\mathcal{F}$ 5% + $\mathcal{P}$ 95%	158.5	33.7
Ours	$\mathcal{F}$ 10% + $\mathcal{P}$ 90%	196.7	35.8
Ours	$\mathcal{F}$ 20% + $\mathcal{P}$ 80%	273.1	37.1
Ours	$\mathcal{F}$ 30% + $\mathcal{P}$ 70%	349.5	38.0
Ours	$\mathcal{F}$ 50% + $\mathcal{P}$ 50%	502.3	38.8
Fully Supervised Models			
MRCNN [8]	$\mathcal{F}$ 100%	884.2	38.8
CondInst [15]	$\mathcal{F}$ 100%	884.2	39.1
SOLOv2 [18]	$\mathcal{F}$ 100%	884.2	39.7

Table 3. **Additional comparisons with weakly-supervised methods in terms of labeling budget and accuracy**. We compare the methods on the COCO *test-dev* under various supervisions; $\mathcal{B}$ (box label), $\mathcal{P}_{10}$ (10-points label), $\mathcal{P}$ (single-point label), $\mathcal{F}$ (full mask label). All methods use the same backbone network of ResNet-101 [9].

egy could give additional performance improvement but demands a more complex training pipeline. In this work, we suffer from the insufficient mask representation of the network when the amount of fully labeled data is extremely limited ($e.g.$, COCO 1%). Although we can alleviate the problem with the proposed MaskRefineNet, we additionally try to adopt this strategy since we assume that the trained student network may have stronger mask representation ability than the teacher network. For this, after training the student network, we newly generate pseudo instance masks for point labeled images. Using both full labels and new pseudo labels, we train a new student network. As the results in table 2, the iterative training strategy yields meaningful improvements on tiny fully labeled data conditions (COCO 1%: 23.9%→25.6%). However, there is no significant performance improvement for subsets above COCO 30%. This result demonstrates that (1) the iterative training strategy is helpful only when the amount of fully labeled data is extremely limited, (2) in more generous conditions such as COCO 30% and 50%, our MaskRefineNet is enough to replenish the mask representation of the network.

Additional Comparison with weakly-supervised method. Point-sup [5] introduced a new type of weak supervision source, multiple (10) points. They achieved remarkable instance segmentation results with a highly reduced annotation cost. To compare with them, we estimate the annotation time for 10-points according to the literature; they labeled 10-points in the bounding box region.

Method	5%	10%	20%	30%	40%	50%
Point DETR [4]	26.2	30.3	33.3	34.8	35.4	35.8
Group R-CNN [21]	30.1	32.6	34.4	35.4	35.9	36.1
ours	32.4	34.3	35.6	36.9	37.0	37.6

Table 4. **Qualitative comparisons on COCO *test-dev* object detection benchmark**. All methods used the ResNet-50 backbone.

- **10 Points**: $77.1 classes/img \times 1s/class + 7.2 inst/img \times (7s/bbox + 10 points \times 0.9s/point) =$ **192.3s/img**.

In table 3, we provide the results for weakly-supervised methods and ours on COCO *test-dev* in terms of accuracy and labeling budget. Although Point-sup shows a slightly better efficiency than ours (37.7% with a budget of 263.2 days vs. 37.1% with a budget of 273.1 days), we argue that our training setting is more applicable for the current dataset conditions than them because they require newly annotating of 10-points. Also, we show the possibility for more performance improvement up to 38.8%, which is highly close to the result of the fully-supervised setting. Furthermore, they give us a new future direction; incorporating 10-points and single-point without any mask labels.

Comparison with weakly semi-supervised object detection methods. In our main paper, we discussed the weakly semi-supervised object detection (WSSOD) methods [4, 21], which used the box labels as strong labels and the point labels as weak labels. Since the instance segmentation covers object detection, we measure our performance on the COCO *test-dev* object detection benchmark. For this, we use the min-max points from the instance mask output as our bounding box output. Even though our strong label is different from theirs (full mask vs. bounding box), the results in table 4 show that ours can surpass the state-of-the-art WSSOD performance. We note that all methods use the same ResNet-50 [9] backbone network and the same amount of total strong and weak labels.

Qualitative analysis for the effect of input sources of MaskRefineNet. In Table 3 of our main paper, we provided the quantitative analysis of the effect of input sources of MaskRefineNet. Here, we supplement our analysis with the qualitative results according to the input sources of the MaskRefineNet in Figure 1. When given all three informative input sources, the MaskRefineNet can produce high-quality refined masks by separating overlapping instances and removing noisy pixels.

Qualitative comparison of baselines and our WSSIS method. In Figure 6 of our main paper, we provided the AP evolution of two baselines and our WSSIS method according to the COCO subsets. In Figure 2, we provide the qualitative results of two baselines and our method under the COCO 10% setting. There are four types of methods: (a) training with fully labeled data only, (b) training with fully labeled data and unlabeled data, (c) training with fully labeled data and point labeled data, and (d) training with fully labeled data and point labeled data along with our point-guided MaskRefineNet. The results demonstrate that the network trained with our method can be guided with higher-quality pseudo labels, resulting less false-positive and false-negative outputs.

Additional qualitative results on COCO dataset. In Figure 3, we provide additional qualitative results of ours trained with 5%, 20%, and 50% COCO subsets.

Qualitative results on BDD100K dataset. We qualitatively analyze the effect of leveraging point labels for the instance segmentation model using the BDD100K dataset [20]. There are two types of networks: the first is the network trained with only 7K fully labeled data, and the second is the network trained with 7K fully labeled data and 67K point labeled data. As shown in Figure 4, due to our effective leveraging of the point labels, the second network is much more robust to large and small instances and occluded instances.

References

[1] Aditya Arun, CV Jawahar, and M Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *European Conference on Computer Vision*, pages 254–270. Springer, 2020. 1

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[3] Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019. 1

[4] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8832, 2021. 2

[5] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2617–2626, 2022. 2

[6] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. *arXiv preprint arXiv:2210.05174*, 2022. 2

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[9] 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. 2

[10] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[11] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1

[12] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 2

[13] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[15] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020. 2

[16] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 2

[17] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1354–1362, 2018. 1

[18] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 2

[19] Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, and Lirong Yang. Centermask: single shot instance segmentation with point representation. In *Proceedings of*

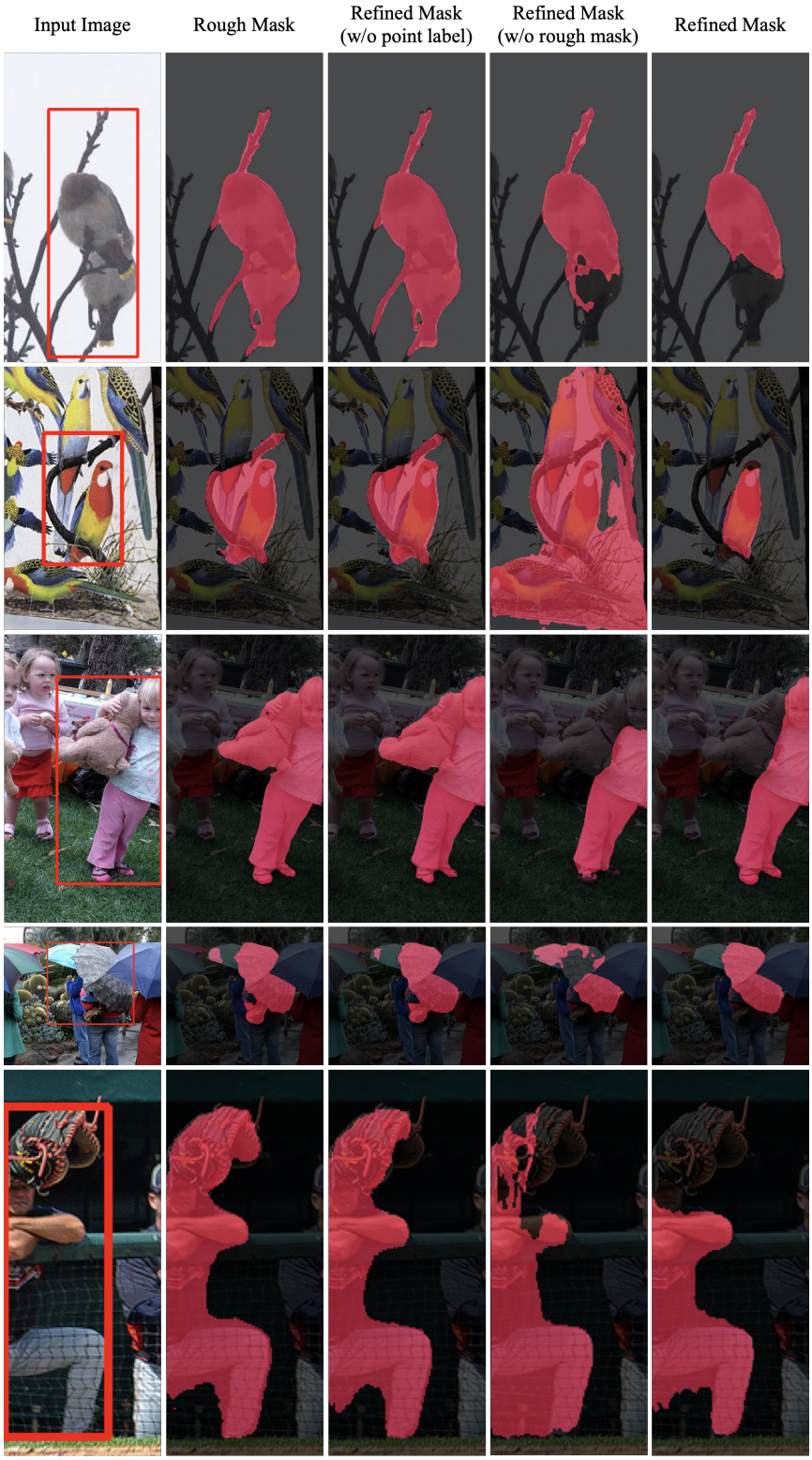

Figure 1. **Qualitative analysis of the effect of input sources of MaskRefineNet**. Object instances can not be distinguished when the point label is not given for MaskRefineNet (3rd col). Meanwhile, mask representations are inaccurate due to the absence of prior rough masks (4th col). Based on these low-cost priors, we can obtain sophisticated masks per object instance (5th col).

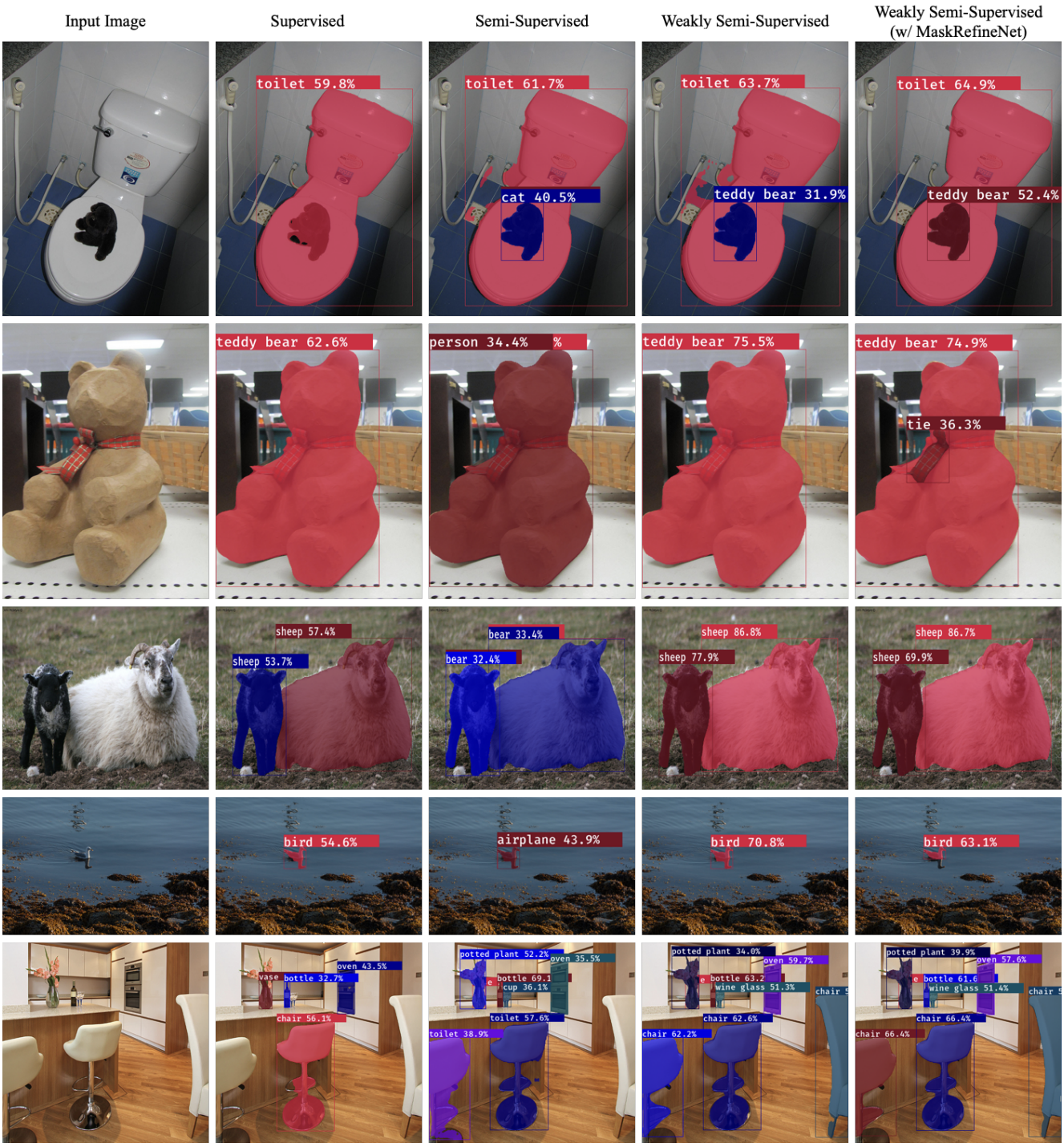

Figure 2. **Qualitative comparison of models trained with different types of supervision on COCO 10% setting**. The result of the semi-supervised setting can detect instance masks for all objects but is vulnerable to misclassification (e.g. cat, person, bear, airplane, toilet). Meanwhile, our point-guided model presents accurate class predictions. Our MaskRefineNet further elaborates the mask representation.

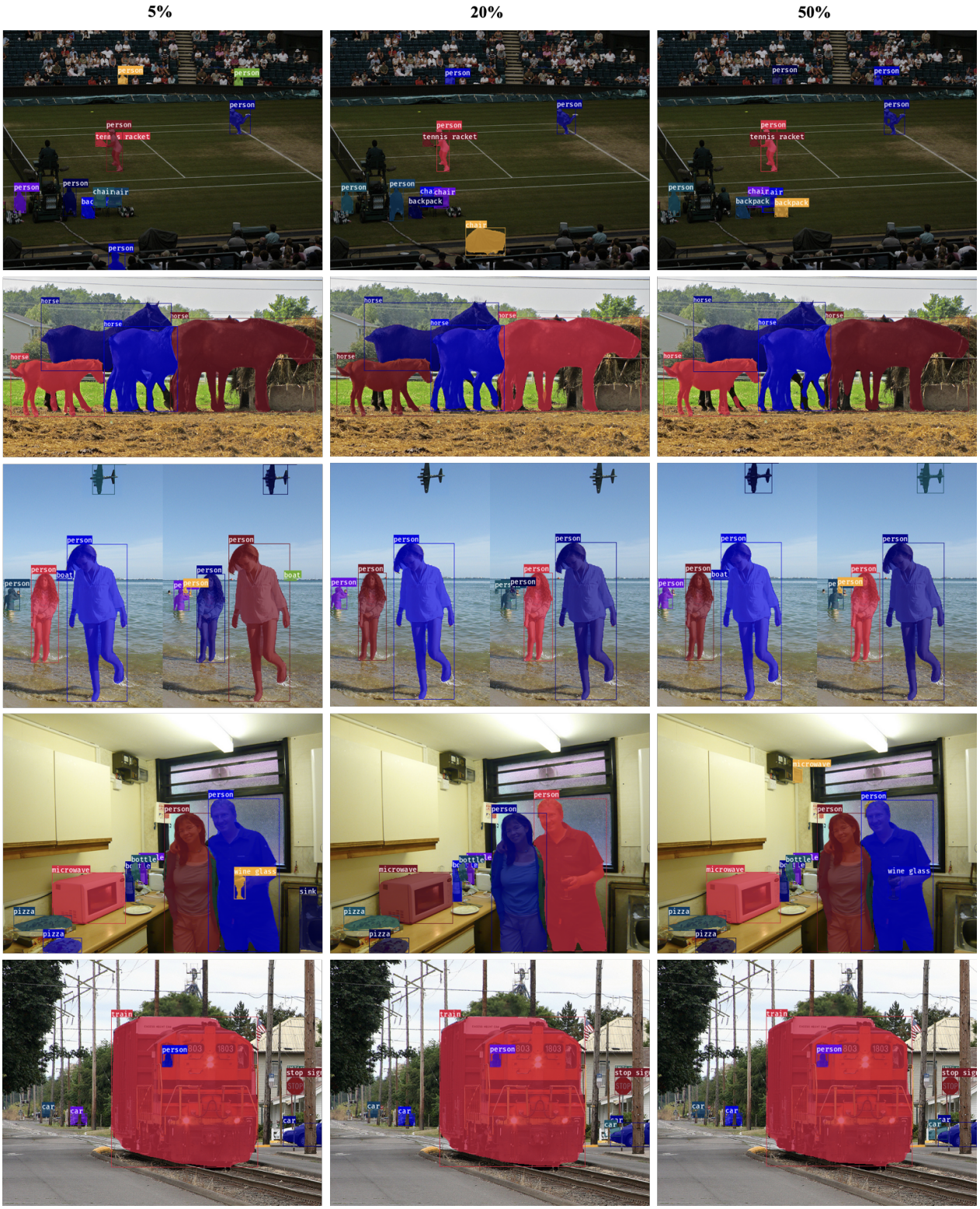

Figure 3. **Additional qualitative results according to the various subsets in COCO data**. Owing to our point guidance along with MaskRefineNet, leveraging only 5% of full labeled data sufficiently localizes all the instances with elaborative mask representations.

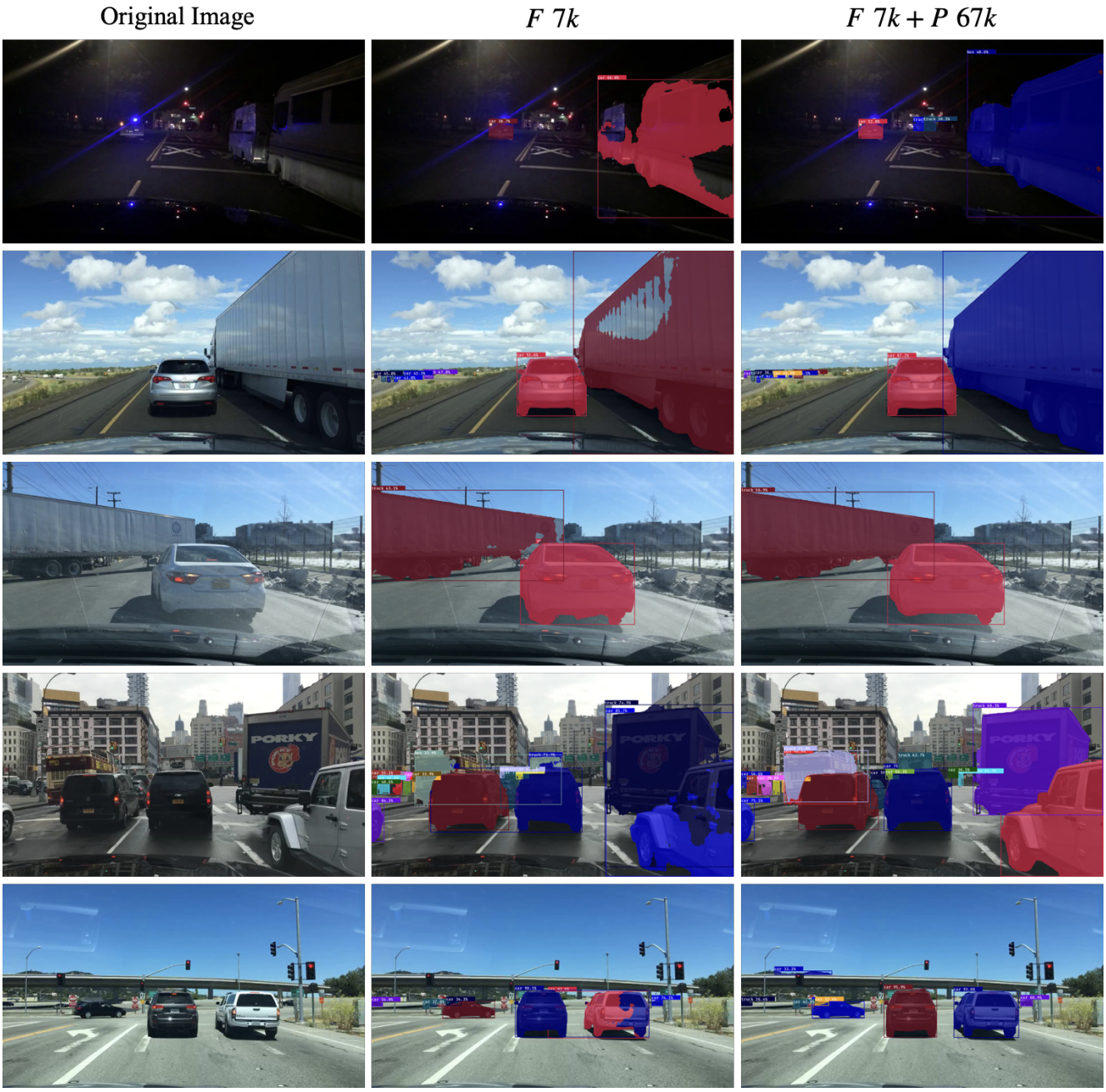

Figure 4. **Qualitative comparison of leveraging point labels on BDD100K**. Training with point labels clearly enriches the mask representation and removes the noise incurred by visually hard samples (*e.g.*, dark light condition in the first row).

the IEEE/CVF conference on computer vision and pattern recognition, pages 9313–9321, 2020. 1

[20] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3

[21] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen. Group r-cnn for weakly semi-supervised object detection with points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9417–9426, 2022. 2

[22] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1