Point2Mask: Point-supervised Panoptic Segmentation via Optimal Transport - *Supplementary Material*

Wentong Li[1], Yuqian Yuan[1], Song Wang[1],
Jianke Zhu[1]*, Jianshu Li[2], Jian Liu[2], Lei Zhang[3]
[1]Zhejiang University [2]Ant Group [3]The HongKong Polytechnical University

A. Sinkhorn Iteration

The transport solver involves the resolution of a linear program in polynomial time. In our OT-based approach, the dimension of pixel samples can be as high as the square of hundreds. To efficiently tackle such a large-scale transport problem, we adopt the Sinkhorn Iteration method [2, 4], which computes the OT problem through the Sinkhorn's matrix scaling algorithm.

The Sinkhorn Iteration converts the OT optimization target into a non-linear but convex form with an entropic regularization term R, which can be formulated as below:

$$\min_{\Gamma_{ij}\in\Gamma} \sum_{i,j=1}^{m,n} \Gamma_{ij} c_{ij} + \lambda R(\Gamma_{ij}), \tag{1}$$

where $R(\Gamma_{ij}) = \Gamma_{ij}(\log \Gamma_{ij} - 1)$, and λ is a regularization coefficient. According to the Sinkhorn-Knopp Iteration method [2, 13], v_i and u_j are introduced for updating the solution:

$$u_j^{t+1} = \frac{y_j}{\sum_i K_{ij} v_i^t}, \quad v_i^{t+1} = \frac{x_i}{\sum_j K_{ij} u_j^{t+1}}, \tag{2}$$

where $K_{ij} = e^{(-c_{ij}/\lambda)}$. After performing the iteration for T times, the optimal plan Γ can be obtained as:

$$\Gamma = \mathbf{diag}(u) K \mathbf{diag}(v). \tag{3}$$

B. Semantic Map Learning

The local LAB affinity and the long-range RGB affinity are integrated to generate the accurate semantic map P^s for the unlabeled regions. In the following, we introduce the two loss terms in detail.

Local LAB Loss. As in [14], the local LAB loss $\mathcal{L}_{sem}^{LAB}$ explores the color similarity $\mathcal{S}_{LAB}$ in LAB color space of the input image with the local kernel. $\mathcal{S}_{LAB}$ is defined as follows:

$$\mathcal{S}_{LAB} = \underset{j\in\mathcal{N}_8(i)}{\mathcal{S}(r_i, r_j)} = \exp\left(-\frac{\|r_i - r_j\|}{\theta_1}\right), \tag{4}$$

where r_i is the LAB color value of pixel i and $\mathcal{N}_8(i)$ denotes its eight local neighbors. θ_1 is the constant parameter. The $\mathcal{L}_{sem}^{LAB}$ loss term is formulated as follows:

$$\mathcal{L}_{sem}^{LAB} = -\frac{1}{z_1} \sum_{i=1}^{n} \sum_{j\in\mathcal{N}_8(i)} \mathbb{1}_{\{\mathcal{S}_{i,j}^{LAB} \geq \tau\}} \log {P_i^s}^T P_j^s, \tag{5}$$

where $z_1 = \sum_{i=1}^{n} \sum_{j\in\mathcal{N}_8(i)} \mathbb{1}_{\{\mathcal{S}_{i,j}^{LAB} \geq \tau\}}$. $\mathbb{1}_{\{\mathcal{S}_{i,j}^{LAB} \geq \tau\}}$ is the indicator function, being 1 if $\mathcal{S}_{i,j}^{LAB} \geq \tau$ and 0 otherwise. As in [14], τ is set to 0.3 and θ_1 is set to 2 by default.

Long-range RGB Loss. Similar to [12], the long-range RGB loss $\mathcal{L}_{sem}^{RGB}$ absorbs the global pixel affinity in RGB space. Each pixel in the input image can be constructed by the global RGB pixel similarity $\mathcal{S}_{RGB}$ through the minimum spanning tree (MST) algorithm. The pixel similarity $\mathcal{S}_{RGB}$ in each tree-connected edge $\mathbb{E}$ is defined as follows:

$$\mathcal{S}_{RGB} = \underset{(l,k)\in\mathbb{E}(i,j)}{\mathcal{S}(r_i, r_j)} = \exp\left(-\frac{\sum \|r_l - r_k\|^2}{\theta_2}\right), \tag{6}$$

where r_i is the RGB pixel value of pixel i. l and k are the adjacent pixels in the tree-connected edge $\mathbb{E}_{i,j}$. Like θ_1, θ_2 is a constant value, which is set to 0.02 by default. The $\mathcal{L}_{RGB}^{sem}$ loss term is defined as:

$$\mathcal{L}_{sem}^{RGB} = -\frac{1}{n} \sum_{i=1}^{n} \left| P_i^s - \frac{1}{z_2} \sum_{\forall j\in\Omega} \mathcal{S}_{i,j}^{RGB} P_j^s \right|, \tag{7}$$

where $z_2 = \sum_j \mathcal{S}_{i,j}^{RGB}$, and Ω denotes the set of pixels in P^s.

C. Additional Results

C.1. Performance on Multiple Point Labels

To further investigate the effectiveness of our approach with multiple point labels, we conduct the experiments with ten-points annotation per target. The results of fully mask-supervised and single point-supervised methods are

*Corresponding author.

Method	Backbone	Supervision	VOC 2012			COCO		
			PQ	PQ^{th}	PQ^{st}	PQ	PQ^{th}	PQ^{st}
Panoptic FPN [5]	ResNet-50	$\mathcal{M}$	65.7	64.5	90.8	41.5	48.3	31.2
Panoptic FCN [10]	ResNet-50	$\mathcal{M}$	67.9	66.6	**92.9**	43.6	49.3	35.0
Panoptic SegFormer [11]	ResNet-50	$\mathcal{M}$	**67.9**	**66.6**	92.7	**48.0**	**52.3**	**41.5**
PSPS [3]	ResNet-50	$\mathcal{P}$	49.8	47.8	89.5	29.3	29.3	29.4
Point2Mask (Ours)	ResNet-50	$\mathcal{P}$	54.2	52.4	90.3	32.4	32.6	32.2
Panoptic FCN [10]	ResNet-50	$\mathcal{P}_{10}$	48.0	46.2	85.2	31.2	35.7	24.3
PSPS [3]	ResNet-50	$\mathcal{P}_{10}$	56.6	54.8	91.4	33.1	33.6	32.2
Point2Mask (Ours)	ResNet-50	$\mathcal{P}_{10}$	59.1	57.5	91.8	35.2	36.1	34.0
Point2Mask (Ours)	ResNet-101	$\mathcal{P}_{10}$	**60.2**	**58.6**	**92.1**	**36.7**	**37.3**	**35.7**

Table A1: Performance comparison on Pascal VOC `val` and COCO `val2017`. $\mathcal{M}$ is pixel-wise mask label. $\mathcal{P}$ and $\mathcal{P}_{10}$ denote 1 and 10 point labels per target, respectively. The results with $\mathcal{M}$ and $\mathcal{P}$ supervision are listed as reference to illustrate the performance with 10 point labels.

Iter. Num.	PQ	PQ^{th}	PQ^{st}
40	53.0	51.2	90.1
60	53.5	51.7	90.1
80	**53.8**	**51.9**	**90.5**
100	52.7	50.8	90.1
120	52.2	50.3	90.2

Table A2: The results with different number of iterations in the Sinkhorn Iteration.

β	PQ	PQ^{th}	PQ^{st}
1.0	52.3	50.4	90.2
0.5	52.4	50.5	90.2
0.2	52.8	50.9	90.3
0.1	**53.8**	**51.9**	**90.5**
0.05	53.1	51.2	90.1
0.01	51.9	50.0	89.6

Table A3: Results with different values of β in Eq. 3 of the main paper.

also listed as reference. As shown in Table A1, we compare Point2Mask with the state-of-the-art methods, including Panoptic FCN [10] and PSPS [3] with ten-points labels on Pascal VOC and COCO datasets. With ResNet-50 backbone, Point2Mask outperforms Panoptic FCN [10] by 11.1% PQ (59.1% *vs.* 48.0%) on Pascal VOC and 4.0% PQ (31.2% *vs.* 35.2%) on COCO. Compared with PSPS [3], Point2Mask surpasses PSPS [3] by 2.5% PQ and 2.1% PQ on Pascal VOC and COCO, respectively. Furthermore, Point2Mask achieves more competitive performance with 60.2% PQ on Pascal VOC and 36.7% PQ on COCO using ResNet-101 backbone.

C.2. Hyper-parameter Selection in OT

We perform the following experiments to examine the impact of hyper-parameters in our proposed OT-based method.

Different Number of Sinkhorn Iterations. We perform Sinkhorn Iteration with different number of iterations to solve the OT problem. Table A2 reports the panoptic segmentation results. When the iteration number is set to 80, Point2Mask achieves the best performance with 53.8% PQ.

Impact of β. In our paper, β in Eq. 3 indicates the importance of boundary map P^b to calculate the pixel-to-gt cost $c_{i,j}$. Table A3 shows the results with different values of β. When $\beta = 0.1$, Point2Mask obtains the best performance.

Figure A1: Visual examples of high-level boundary map. The accurate boundary for thing-based objects can be learnt.

This indicates that the cost from instance-wise boundary map P^b plays a complementary role to the main cost term based on the category-wise semantic map P^s. Furthermore, the visual examples of learnt high-level boundary P^b_{high} are shown in Fig. A1.

C.3. More Visualization Results

To further illustrate the performance of our single point-supervised approach, we give more visualization results.

Fig. A2 shows the qualitative comparison with the state-of-the-art method PSPS [3]. It can be seen that our proposed Point2Mask approach is able to find the ambiguous locations of nearby instances precisely. This demonstrates that our OT-based approach can discriminate the thing-based targets with the accurate boundaries. In addition, Fig. A3

provides the panoptic segmentation results of Point2Mask on general COCO and Pascal VOC datasets.

D. Discussion

Differences against the existing works. Like previous weakly-supervised methods [3, 14, 9, 8], our method aims to achieve high-quality segmentation with the label-efficient sparse labels, which is different from the existing promptable segmentation model [6] with a large amount of data and the corresponding mask labels.

We adopt the same base architecture as PSPS [3], *i.e.*, generating pseudo labels firstly and then training the panoptic segmentation branch. To generate the panoptic pseudo labels, both our method and PSPS [3] employ the category-wise and instance-wise representations. For category-wise representation, we firstly employ the local LAB and long-range RGB pixel similarities (Sec.3.4.1), instead of the local LAB semantic parsing only as in [3]. Secondly, for instance-wise representation, we adopt the boundary map and define different distance functions. Compared with the high-level manifold cues in [3], the boundary map is more suitable for the shortest path-based implementation to calculate the instance-wise differences. More importantly, *the key difference lies in the presented OT formulation for global assignment to generate more accurate mask labels*.

Limitations. For the dense objects with the same categories, such as in autonomous driving and remote sensing scenarios, the proposed method may not perform well with the supervision of only a single point label. Better performance can be obtained by adopting the more powerful segmentation network, like Mask2Former [1] and MaskDINO [7], into our method.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1290–1299, 2022. 3

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 26, 2013. 1

[3] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 319–336. Springer, 2022. 2, 3

[4] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 303–312, 2021. 1

[5] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6399–6408, 2019. 2

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[7] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2023. 3

[8] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *Proc. Eur. Conf. Comp. Vis.*, pages 1–18. Springer, 2022. 3

[9] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu, Xiansheng Hua, and Lei Zhang. Box2mask: Box-supervised instance segmentation via level-set evolution. *arXiv preprint arXiv:2212.01579*, 2022. 3

[10] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 2

[11] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1280–1289, 2022. 2

[12] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 16907–16916, 2022. 1

[13] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 1

[14] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5443–5452, 2021. 1, 3

Figure A2: Qualitative comparisons on Pascal VOC. The left two columns show that Point2Mask can precisely discriminate the nearby instances of the same category. The right two columns indicate that Point2Mask can obtain more fine-grained boundaries.

Figure A3: Visual examples of panoptic segmentation by our Point2Mask with single point label per target on COCO and Pascal VOC datasets.