Supplementary Material

Rui Hou[*,1,2] Jie Li[*,1] Arjun Bhargava[1] Allan Raventos[1] Vitor Guizilini[1] Chao Fang[1]
Jerome Lynch[2] Adrien Gaidon[1]

[1]Toyota Research Institute [2]University of Michigan, Ann Arbor
[1]{firstname.lastname}@tri.global [2]{rayhou, jerlynch}@umich.edu

We provide more detailed descriptions, metrics, and visualizations for our proposed approach that were not included in the main text due to space limitation:

- per-class metrics of our model on Cityscapes and COCO (Tables 1 and 2 in the main text);
- implementation details for the comparison without levelness presented in the ablative analysis (Table 3 in the main text);
- detailed description and discussion of the proposed weakly supervised application of our method (Section 5.6 Table 3 in the main text);

1. Per-class Performance

We provide per-class PQ metrics of our models on Cityscapes and COCO dataset in Table S1 and Table S2, which corresponds respectively to the entries in Table 1 and Table 2 in the main text. For both datasets, $\sigma = 0.3$ is used as the foreground mask acceptance probability in Eq. 8, which is determined through hyper-parameter grid search.

2. Contribution of Levelness Map

As we have shown in the paper (also depicted in Figure 3), our proposed model predicts a global levelness map using feature maps produced by the localization tower at all scale levels. This levelness map indicates which scale level does the bounding box at each location (x, y) belongs to (with 0 reserved for background). At inference time, the levelness map is used to provide indexes while we assemble the global dense bounding box prediction $\mathcal{B}(x, y)$ from each FPN level ($\{\mathbf{b}^i_{xy}\}$), according to Eq. 15.

However, the levelness is not a necessity in our proposed model, as the assembling process can be done without it. In our ablative analysis in Section 5.5, we compared our method to a simple alternative assembling approach without levelness to justify the value of such design.

This alternative solution, instead of assembling a unique bounding box for each feature location (x, y), carries all the bounding box predictions from different FPN levels i to the foreground mask probability estimation:

$$\mathcal{B}(x, y; i) = \mathbf{b}^i_{xy} \tag{S1}$$

Then the location based foreground probability becomes:

$$\hat{P}_{loc}(x, y, j; i) = \text{IoU}(\mathcal{B}(x, y; i), \mathbf{B}_j) \tag{S2}$$

We can still obtain a single probability for each location by taking the maximum along the level dimension i:

$$\hat{P}_{loc}(x, y, j) = \underset{i}{\text{argmax}}(\hat{P}_{loc}(x, y, j)) \tag{S3}$$

We report the resulting model performance of this alternative solution in the second entry of the ablative analysis table (Table 3 in the main text). For simplicity, this comparison on levelness is done without the mask loss. All other configurations are the same as the default model. This comparison indicates that the levelness map, with only 1 additional convolutional layer, is able to provide a better cross-level indication that results in an increase in performance (+1% PQ).

3. Weakly Supervised Application

In the weakly supervised scenario discussed in Section 5.6 of the main text, we consider relying only on semantic and bounding box labels for the panoptic segmentation task.

Figure S1: **Weak Supervision** Left: input image; Middle: fully supervised pixel association; Right: Weakly supervised pixel association. In this example, orange, blue and yellow color indicates the pixels which are assigned to a bounding box target of the three cars during training.

These two types of labels present a weaker supervision for the panoptic segmentation task, as there are ambiguous

Class	PQ	SQ	RQ	Class	PQ	SQ	RQ
Mean	58.81	79.81	72.32	road	97.58	97.88	99.69
sidewalk	75.91	83.87	90.51	building	87.67	89.49	97.96
wall	30.02	72.21	41.57	fence	34.89	73.27	47.62
pole	50.15	65.23	76.88	traffic light	45.98	70.97	64.79
traffic sign	68.23	77.32	88.24	vegetation	88.78	90.27	98.35
terrain	34.45	73.75	46.72	sky	86.73	92.16	94.10
person	47.96	75.76	63.30	rider	49.61	70.76	70.11
car	60.55	83.30	72.69	truck	47.16	83.12	56.74
bus	68.51	88.75	77.19	train	55.94	83.91	66.67
motorcycle	44.54	72.68	61.28	bicycle	42.86	71.76	59.72

Table S1: **Per-class Performance on Cityscapes**

pixels between overlapping bounding boxes of the same category.

In our proposed algorithm, we directly regress the bounding box and semantic classification. The construction of instance masks relies on the accuracy of the bounding box predictions instead of the explicit modeling of instance shapes. Thus, our method is robust to the absence of foreground mask information. In our implementation, we relax the pixel assignment during bounding box prediction as depicted in Figure S1, favoring the smallest bounding box in overlapping cases.

As shown in Table 3, our weakly supervised model achieved 55.7 PQ, i.e. 95% of the performance of the fully supervised model. Some qualitative examples are also provided in Figure S2.

Figure S2: **Weakly supervised panoptic segmentation** Our proposed algorithm obtain promising and practical prediction results, trained with only bounding box and semantic labels.

Class	PQ	SQ	RQ	Class	PQ	SQ	RQ
Mean	37.13	76.14	46.98	tv	61.49	84.71	72.59
bed	52.43	83.47	62.82	bus	66.56	86.39	77.05
car	45.06	76.82	58.66	cat	71.66	86.64	82.71
cow	53.58	77.06	69.53	cup	43.82	81.58	53.72
dog	62.20	83.26	74.70	net	40.53	77.81	52.08
sea	72.62	89.61	81.03	tie	23.70	69.93	33.89
bear	73.45	85.47	85.94	bird	34.53	75.04	46.01
boat	28.71	69.99	41.02	book	12.43	66.94	18.57
bowl	39.12	79.67	49.10	cake	41.43	80.86	51.23
fork	13.32	67.63	19.70	kite	33.84	71.18	47.54
oven	47.03	80.87	58.15	road	52.09	81.22	64.14
roof	12.68	66.09	19.19	sand	50.00	85.77	58.29
sink	46.48	79.31	58.60	skis	4.17	59.49	7.02
snow	79.12	90.81	87.12	tent	8.00	68.01	11.76
vase	40.49	76.72	52.77	apple	24.08	77.40	31.11
bench	21.75	75.45	28.83	chair	29.38	73.59	39.92
clock	58.55	82.25	71.19	couch	46.16	83.70	55.16
donut	47.41	84.45	56.13	fruit	4.42	56.74	7.79
horse	55.45	77.40	71.64	house	17.33	68.41	25.33
knife	7.36	71.53	10.28	light	15.48	67.57	22.91
mouse	59.03	82.46	71.58	pizza	55.86	84.97	65.74
river	43.97	86.54	50.81	sheep	48.27	76.27	63.29
shelf	13.45	62.14	21.65	spoon	3.95	70.62	5.59
towel	21.27	75.28	28.25	train	69.25	87.58	79.07
truck	37.23	79.59	46.78	zebra	62.10	79.77	77.85
banana	23.25	74.59	31.17	banner	12.18	72.80	16.72
bottle	40.52	76.06	53.27	bridge	12.35	62.64	19.72
carrot	20.33	70.37	28.89	flower	17.92	78.22	22.92
gravel	13.75	70.55	19.49	laptop	52.07	80.26	64.88
orange	29.22	81.95	35.66	person	55.63	76.64	72.58
pillow	1.34	67.86	1.98	remote	22.41	74.18	30.22
stairs	12.77	68.65	18.60	toilet	68.16	86.37	78.92
bicycle	31.45	70.83	44.40	blanket	4.92	66.76	7.37
counter	18.58	69.66	26.67	curtain	42.60	78.10	54.55
frisbee	57.23	81.01	70.65	giraffe	63.30	78.75	80.38
handbag	15.24	72.38	21.05	hot dog	32.36	82.80	39.08
toaster	19.90	64.69	30.77	airplane	60.55	80.01	75.68
backpack	18.91	75.04	25.20	broccoli	29.18	72.30	40.35
elephant	65.76	80.56	81.62	keyboard	49.18	80.80	60.87
platform	19.08	79.62	23.96	railroad	44.68	72.61	61.54
sandwich	33.86	81.97	41.31	scissors	31.51	75.92	41.51
suitcase	42.38	78.90	53.71	umbrella	51.68	79.63	64.90
cardboard	16.68	68.63	24.31	microwave	57.83	82.50	70.10
snowboard	19.16	72.23	26.53	stop sign	67.28	91.11	73.85
surfboard	39.52	74.85	52.80	wall-tile	44.76	77.36	57.86
wall-wood	19.96	72.12	27.67	cell phone	32.06	80.52	39.81
door-stuff	20.98	72.71	28.85	floor-wood	43.09	79.43	54.25
hair drier	0.00	0.00	0.00	motorcycle	47.36	76.38	62.01
rug	37.32	78.32	47.65	skateboard	44.63	71.17	62.71
teddy bear	52.60	81.06	64.90	toothbrush	10.08	69.57	14.49
wall-brick	27.88	72.79	38.31	wall-stone	17.45	76.97	22.67
wine glass	35.82	76.88	46.59	dirt	30.02	77.54	38.72
rock	32.14	77.30	41.57	sports ball	45.27	78.09	57.97
tree	64.41	80.82	79.70	water-other	22.09	81.91	26.97
baseball bat	23.83	66.73	35.71	dining table	29.18	74.86	38.98
fence	26.39	71.53	36.89	fire hydrant	65.68	83.51	78.65
grass	54.96	82.77	66.40	mirror-stuff	26.98	73.91	36.50
paper	11.18	67.05	16.67	playingfield	62.58	88.66	70.59
potted plant	28.78	71.20	40.42	refrigerator	57.87	85.28	67.86
table	20.75	72.09	28.79	window-blind	37.01	79.52	46.54
window-other	27.50	72.53	37.92	parking meter	54.27	81.41	66.67
tennis racket	57.19	78.94	72.45	traffic light	38.18	74.12	51.51
baseball glove	36.60	76.41	47.90	cabinet	40.15	77.20	52.01
ceiling	50.09	78.23	64.03	mountain	41.87	75.86	55.20
pavement	36.49	77.79	46.90	sky-other	81.65	90.76	89.96
food-other	13.83	72.99	18.95	wall-other	44.06	77.33	56.97
floor-other	38.16	78.39	48.68	building-other	38.17	77.90	49.00

Table S2: **Per-class Performance on COCO**