Supplementary Materials to
"MaSS13K: A Matting-level Semantic Segmentation Benchmark"

In this supplementary file, we provide the following materials:

- More samples from our MaSS13K dataset (referring to Sec. 3 of the main paper).
- More details on training and evaluation metrics (referring to Sec. 5 of the main paper).
- More experimental results (referring to Sec. 5.1, 5.2, 5.3 of the main paper).
- Limitations.

A. MaSS13K Dataset

A.1. More Samples from MaSS13K

We provide more annotated samples from our MaSS13K in Fig. 1. MaSS13K covers a diverse range of indoor and outdoor scenes, such as urban areas, natural landscapes, street views, wilderness, parks, mall interiors, and other public spaces.

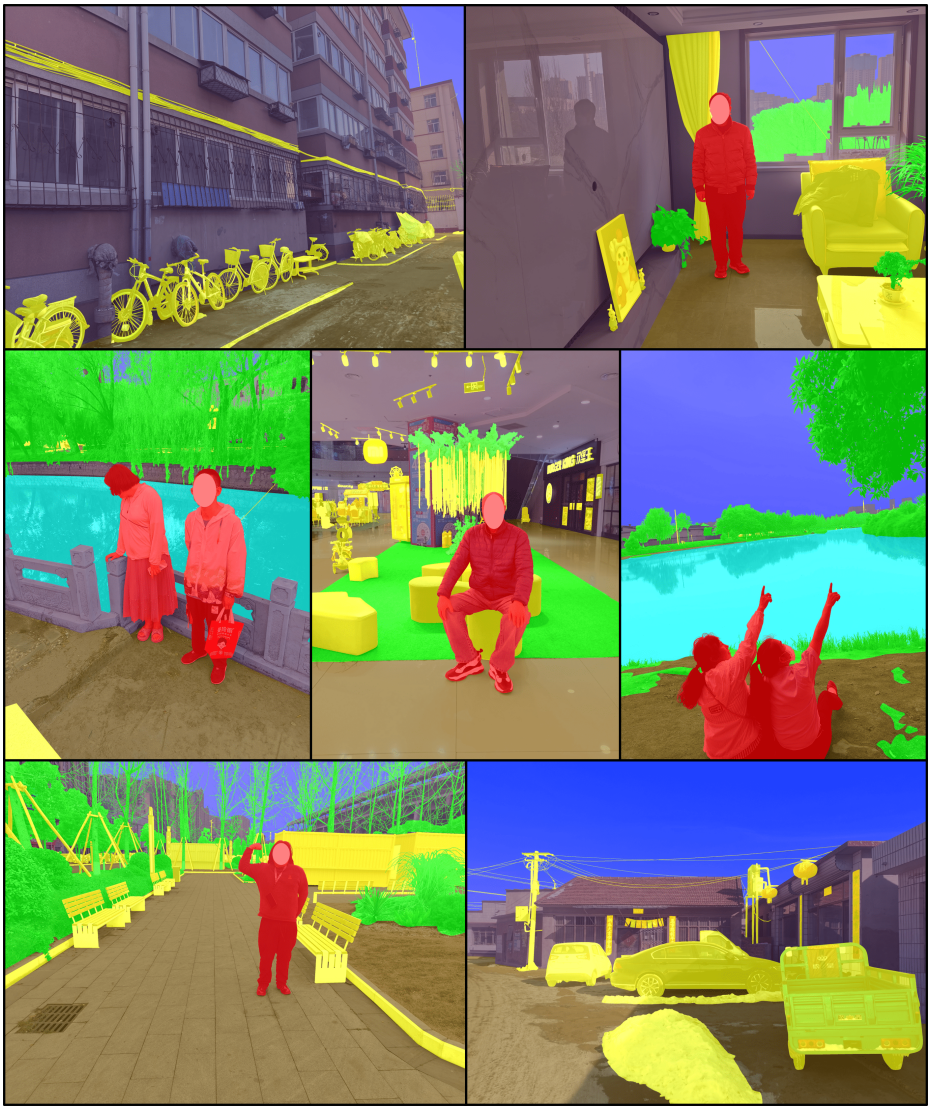

Figure 1. More samples from our MaSS13K dataset.

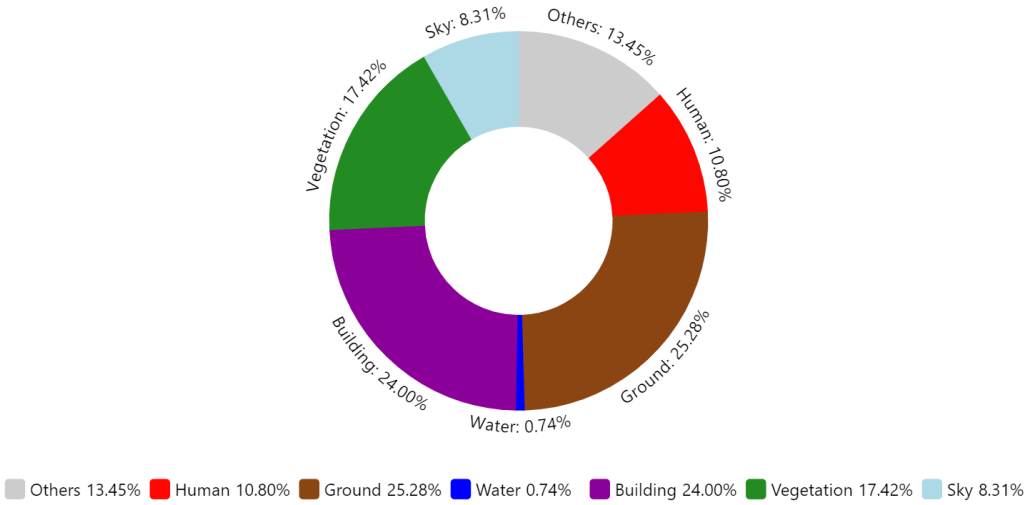

Figure 2. Pixel distribution of the seven categories in MaSS13K.

A.2. Pixel Distribution of the Categories in MaSS13K

Fig. 2 illustrates the pixel distribution of the seven annotated categories in MaSS13K. Except for the "`Water`" category, the pixel distribution of the other categories is relatively balanced. The dataset primarily consists of outdoor street scenes and urban landscapes, resulting in the highest pixel proportions for "`Ground`" and "`Building`". Additionally, 13.45% of the pixels in the dataset are classified as "`Others`", which includes various objects such as traffic signs, billboards, and vehicles. Despite being grouped into the "`Others`" category, these objects are accurately labeled, ensuring the high quality of the dataset and its potential for further processing and development.

B. More Details on Training and Evaluation Metrics

B.1. Training Details

We implement MaSSFormer using mmsegmentation toolbox [3]. We use the AdamW [7] optimizer to train our model with a batch size of 16. The initial learning rate is set to 0.003 with a weight decay of 0.05, and the cosine decay schedule is applied during training. For data augmentation, we mainly follow the setup of Mask2Former [1], including random resizing, random cropping, and random flipping. The models are trained for 80k iterations on the MaSS13K dataset.

For the other evaluated methods on the MaSS13K dataset, we use their default settings on learning rate and data augmentation. The resolution during training and testing for all methods is kept consistent to ensure a fair comparison. The number of parameters and FLOPs for all methods are calculated using the mmsegmentation tools, except for PEM and MPFormer, which are calculated using Detectron2 tools.

B.2. Details of Evaluation Metrics

In high-resolution semantic segmentation, there are numerous fine-grained regions of object details that are critical for the quality of the segmentation masks. However, the standard mask IoU metric is too coarse to differentiate these fine-grained regions, making it less effective in evaluating high-resolution semantic segmentation performance. To better evaluate and compare the methods of high-resolution semantic segmentation, we also use boundary-focused metrics, including Boundary IoU (BIoU) [2] and Boundary F-1 Score (BF1) [4], in the main paper.

BIoU. For the mask of the i-th category, the BIoU^i is defined as follows:

$$\mathrm{BIoU}^i = \frac{(G_d^i \cap G^i) \cap (P_d^i \cap P^i)}{(G_d^i \cap G^i) \cup (P_d^i \cap P^i)}, \tag{1}$$

where P and G denote the predicted and ground-truth maps, respectively, and the subscript d denotes the mask obtained by dilating the boundary by d pixels. In our benchmark, we set d to 0.1% of the diagonal length, which is 5 pixels, to better measure the accuracy of details.

BF1. BF1 Score is a commonly used evaluation metric for edge detection and segmentation that combines precision and recall to assess the edge quality of a segmentation map. The BF1 score is calculated as follows:

$$\mathrm{BF1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{2}$$

where $\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ and $\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. In this case, if a point of predicted boundary matches a ground truth boundary point within a distance error tolerance, it is considered a true positive. In our evaluation, we set the distance error tolerance to 2. "$\mathrm{TP} + \mathrm{FP}$" can be represented by the total number of edge points in the ground truth map, and "$\mathrm{TP} + \mathrm{FN}$" can be represented by the total number of edge points in prediction map.

C. More Experimental Results

C.1. Ablation on Different Backbones

In Tab. 1, we compare the performance of Mask2Former and MaSSFormer using stronger backbones. We can observe that with stronger backbones, the IoU scores become higher due to the improved global semantic extraction capabilities. However, the BIoU and BF1 gaps between Mask2Former and our proposed MaSSFormer are enlarged, further validating that MaSSFormer can better process detail regions and produce higher-quality segmentation edges.

Methods	Backbone	**MaSS-val**			**MaSS-test**			Stat.	
		mIoU	BIoU	BF1	mIoU	BIoU	BF1	Para.	FLOPS
Mask2Former	Swin-T	89.17	48.27	0.5529	89.47	47.97	0.5509	47.40M	3148G
MaSSFormer	Swin-T	89.49 (**+0.32**)	49.68 (**+1.41**)	0.5719 (**+0.0190**)	89.40 (**-0.07**)	49.50 (**+1.53**)	0.5685 (**+0.0176**)	40.92M	1956G
Mask2Former	Swin-B	91.30	51.20	0.5912	91.10	50.83	0.5893	107M	6126G
MaSSFormer	Swin-B	91.35 (**+0.05**)	53.00 (**+1.80**)	0.6102 (**+0.0190**)	91.31 (**+0.21**)	52.74 (**+1.91**)	0.6076 (**+0.0183**)	100M	4890G

Table 1. Quantitative comparison with different backbones.

C.2. Results of Each Category

In Tab. 2 and Tab. 3, we present detailed quantitative results for each category in **MaSS-val** and **MaSS-test**, respectively. From the two tables, it is evident that the '`Others`' class has the lowest IoU scores. This is because the '`Others`' class aggregates a variety of objects beyond the six categories, making it more challenging to identify and leading to lower overall IoU. In contrast, the '`Human`' and '`Sky`' categories are relatively distinct and well-defined, resulting in higher IoU scores. Additionally, we observe that MaSSFormer achieves the most significant improvement in BIoU for the '`Vegetation`' category compared to other methods. This aligns with our visual observations, as trees in the dataset contain a large amount of boundaries and details.

C.3. More Details and Results on Segmenting New Classes

Details of the Pipeline. We utilize the Grounded-SAM [8] as a semi-automatic labeling tool to obtain pseudo labels without human effort. Grounded-SAM employs the bounding boxes from Grounding-DINO [6] as the input to SAM [5] to generate masks for the corresponding objects. Grounding-DINO, as an open-vocabulary detection model, can take any text prompt and produce bounding boxes for objects of that category in the image. Therefore, for a new category '`X`', we input '`X`' as a text-prompt into Grounded-SAM, process all images in the dataset to obtain instance segmentation masks for that category, and then merge and convert these masks into semantic segmentation labels for the new category '`X`'. We combine the semantic labels of the new category with the existing seven categories' labels for joint model training. To provide a quantitative evaluation of the mask quality for the new category, we manually annotate a set of images of that category from the test set to calculate the mIoU, BIoU, and BF1 scores.

More Results. In the main paper, we have validated the effectiveness of our method in segmenting a new class '`Car`' on MaSS13K. In this supplementary file, we conduct experiments on another new class '`Bicycle`'. The results are shown in Tab. 4. For the convenience of readers, we also put the results of class '`Car`' in the table. We can see that for both of the two new classes, the IoU, BIoU, and BF1 metrics show significant improvements over the baseline. Some visual examples

Table 2. Quantitative evaluation on MaSS-val for each category.

Methods	Others			Human			Building			Vegetation			Ground			Sky			Water		
	mIoU	BIoU	BF1	mIoU	BIoU	BF1	mIoU	BIoU	BF1	mIoU	BIoU	BF1	mIoU	BIoU	BF1	mIoU	BIoU	BF1	mIoU	BIoU	BF1
STDC2	66.52	18.39	.2041	95.77	45.06	.5366	77.96	19.78	.2467	87.88	22.84	.2701	90.62	32.54	.4311	89.96	34.21	.3647	73.51	23.66	.2582
BiSeNetV2	53.26	13.31	1589	92.38	38.90	.5087	69.93	15.27	.2013	83.62	27.45	.3013	85.16	26.84	.3940	82.90	41.45	.4637	35.83	11.40	.1417
SegNext	70.17	25.37	.3050	97.58	59.61	.7061	82.22	27.68	3440	89.30	36.58	.4132	93.74	40.94	.5392	92.40	48.89	.5783	88.60	40.50	.4462
PIDNet-L	64.71	19.83	.2258	96.16	44.74	.4597	77.74	21.82	.2620	86.64	29.74	.3278	89.84	32.24	.4324	89.64	41.85	.4515	71.15	28.90	.3144
FeedFormer	65.90	23.34	.2870	97.02	57.68	.6986	79.46	26.97	.3465	89.67	44.59	.5006	92.66	38.93	.5187	94.10	62.19	.7321	91.42	40.73	.4464
SeaFormer	69.75	24.53	.2918	97.10	57.88	.7029	82.63	26.93	.3349	88.41	34.61	.3955	92.16	39.69	.5211	93.59	48.56	.5565	83.83	38.13	.4279
CGRSeg	59.32	21.83	.2618	90.42	47.90	.6129	74.05	23.42	.2917	88.15	34.30	.3843	87.47	34.53	.4699	88.88	46.60	.5486	80.80	32.78	.4031
DeepLabV3+	67.61	22.76	.2947	96.30	56.36	.7074	78.83	25.30	.3359	89.35	41.46	.4831	90.25	34.66	.4935	94.09	60.35	.6822	90.28	40.38	.4518
UperNet	62.60	20.15	.2554	92.60	45.40	.5772	77.32	23.63	.2941	88.20	39.85	.4479	88.79	33.03	.4553	91.44	58.36	.6253	73.46	31.39	3456
OCRNet	61.11	18.34	.2304	93.91	46.63	.6051	75.17	20.19	.2617	88.22	31.15	.3577	89.35	31.17	.4398	92.45	47.05	.5077	87.78	34.61	.3548
MaskFormer	60.16	19.82	.2680	94.55	51.09	.6463	72.94	22.80	.3043	87.83	43.98	.4844	85.69	29.78	.4303	93.96	62.15	.7159	87.72	40.65	.4194
Mask2Former	72.45	31.23	.3907	98.13	66.08	.7755	82.95	32.86	.4165	90.25	45.99	.5170	93.23	44.72	.5824	94.52	63.59	.7417	86.42	47.35	.5241
MPFormer	73.34	31.84	.3913	98.49	67.49	.7885	84.29	33.93	.4258	89.93	46.90	.5305	93.04	44.68	.5772	93.59	64.29	.7475	81.68	45.54	.5139
PEM	68.10	25.46	.3198	97.50	60.32	.7288	70.49	23.61	.3232	89.75	44.25	.4929	77.71	26.98	.4225	93.31	61.66	.7171	87.02	41.30	.4480
MaSSFormer-Lite	67.62	26.02	.3288	96.98	61.01	.7352	79.65	30.01	.3854	88.90	49.31	.5374	90.59	38.90	.5327	95.35	68.26	.7587	90.65	43.97	.4784
MaSSFormer	71.27	30.84	.3931	97.44	64.62	.7783	82.74	33.74	.4385	89.76	50.36	.5799	93.40	44.13	.5605	95.90	69.91	.8108	92.32	49.18	.5594

Table 3. Quantitative evaluation on MaSS-test for each category.

Methods	Others			Human			Building			Vegetation			Ground			Sky			Water		
	mIoU	BIoU	mIoU	BF1	BIoU	mIoU	BF1	BIoU	mIoU	BF1	BIoU	mIoU	BF1	BIoU	mIoU	BF1	BIoU	mIoU	BF1	BIoU	BF1
STDC2	66.75	17.88	.2008	94.98	43.71	.5246	78.97	19.69	.2492	87.72	22.36	.2680	90.35	31.82	.4302	91.20	34.27	.3712	76.37	24.29	.2622
BiSeNetV2	52.63	12.84	.1600	90.53	37.35	.4967	70.37	15.24	.2010	83.31	26.91	.2979	84.90	26.45	.3934	86.50	41.27	.4625	42.23	11.27	.1500
SegNext	69.80	24.63	.3031	97.41	58.45	.6935	82.40	27.45	.3472	89.41	36.24	.4076	93.49	39.78	.5316	93.50	49.49	.5775	90.76	40.12	.4490
PIDNet-L	64.63	19.65	.2269	95.47	43.50	.4505	78.24	22.13	.2668	87.07	29.38	.3210	89.69	31.77	.4304	89.95	42.08	.4586	67.31	26.38	.3191
FeedFormer	65.67	22.47	.2823	96.29	55.91	.6824	79.82	26.70	.3464	89.55	44.06	.4898	92.34	37.80	.5138	94.29	62.28	.7281	88.02	38.31	.4457
SeaFormer	69.30	24.23	.2927	96.23	56.49	.6923	82.68	27.04	.3373	88.66	34.30	.3883	92.13	38.77	.5172	93.61	48.96	.5556	88.96	38.23	.4354
CGRSeg	59.17	21.09	.2596	90.65	46.80	.6038	73.96	23.37	.2952	87.97	33.82	.3778	87.68	33.78	.4686	90.64	46.92	.5426	80.09	31.92	.3942
DeepLabV3+	67.32	22.46	.2930	95.65	54.11	.6906	79.25	25.66	.3363	89.35	41.53	.4763	89.76	33.95	.4859	93.37	60.16	.6816	81.25	32.65	.4224
UperNet	61.97	19.78	.2515	90.70	43.30	.5601	77.74	24.19	.3000	88.20	40.29	.4429	88.20	32.56	.4500	92.19	59.08	.6318	74.87	30.06	.3617
OCRNet	60.60	17.57	.2222	92.41	43.44	.5875	75.76	20.08	.2580	87.74	30.36	.3437	88.48	29.83	.4306	92.65	46.95	.5070	83.56	31.10	.3501
MaskFormer	61.28	19.65	.2679	94.48	50.26	.6430	74.64	22.99	.3059	88.24	44.14	.4818	85.61	29.04	.4241	94.25	62.80	.7160	84.06	36.45	.4182
Mask2Former	70.67	29.60	.3708	97.69	65.02	.7656	80.83	31.94	.4033	90.25	44.37	.4985	91.66	42.60	.5608	94.88	64.13	.7241	90.00	45.27	.5233
MPFormer	72.38	31.23	.3896	98.06	66.13	.7784	83.91	33.67	.4271	90.17	46.70	.5228	92.58	43.68	.5718	93.57	64.81	.7451	79.63	44.02	.5124
PEM	67.70	25.04	.3210	97.46	59.97	.7205	72.14	24.02	.3287	89.12	44.01	.4854	78.75	26.12	.4146	91.86	61.25	.7094	86.69	39.51	.4593
MaSSFormer-Lite	66.80	25.11	.3260	96.39	59.19	.7216	79.60	29.55	.3844	88.77	46.75	.5280	90.87	37.69	.5170	94.77	64.54	.7548	85.71	40.08	.4707
MaSSFormer	70.07	29.70	.3860	97.23	64.02	.7665	81.65	33.40	.4387	90.03	51.91	.5772	91.51	41.56	.5557	95.60	70.30	.8043	90.18	47.82	.5611

Table 4. Quantitative evaluation on novel classes **`Bicycle`** and **`Car`**.

Settings	**`Bicycle`**			**`Car`**		
	mIoU	BIoU	BF1	mIoU	BIoU	BF1
Pseudo label generated by Grounded-SAM	49.40	23.82	0.2800	94.18	20.44	0.2522
Prediction generated by our joint-trained model	74.57	40.17	0.4609	95.21	35.68	0.3643

of the segmentation results of '`Bicycle`' are presented in Fig. 3. We can see that due to the relatively small size of bicycle targets, the incorrect segmentation of the wheels can severely impact the IoU scores, as shown in the 2nd row of Fig. 3. In addition, the failure of Grounded-SAM to detect small targets can further reduce the IoU, as illustrated in the 1st row of Fig. 3. Our method, designed for high-resolution images, can effectively capture fine structural details and boundaries, resulting in higher IoU and BF1 scores. Furthermore, under the joint supervision of other precise categories, our method can accurately distinguish foreground from background at the wheel areas, resulting in precise segmentation of the target objects.

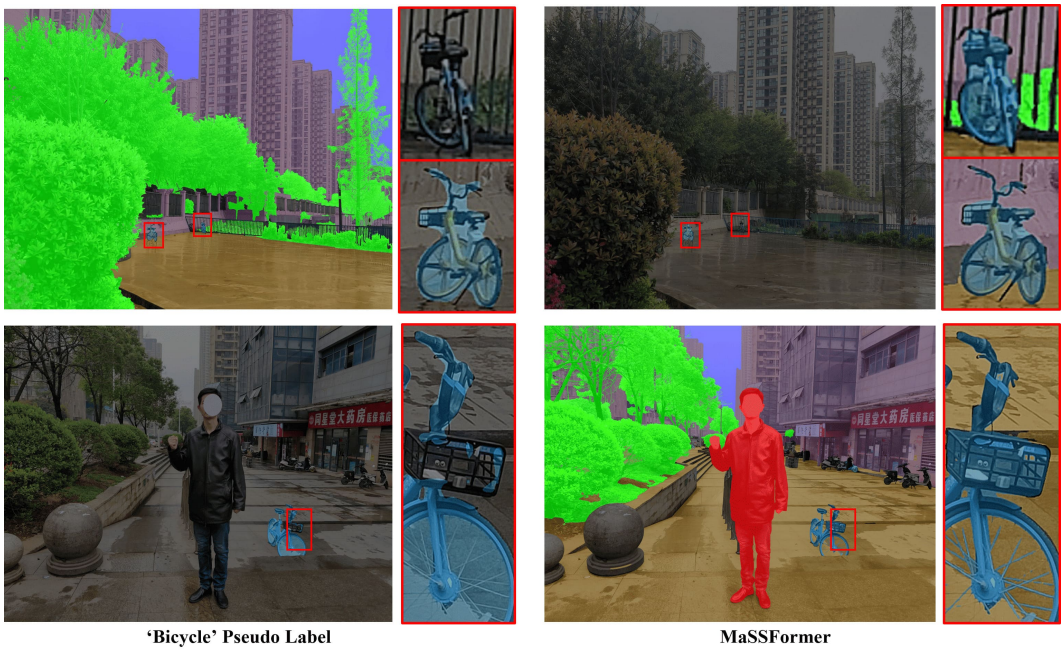

Figure 3. Visual results on the segmentation of '`Bicycle`' class. Left: Pseudo-labels generated by Grounded-SAM. Right: Predictions by MaSSFormer.

C.4. More Qualitative Comparisons

We present more qualitative comparisons between our MaSSFormer and other representative methods in Fig. 4. It can be seen that MaSSFormer demonstrates superior performance in segmenting fine-grained regions, such as the thin lines in the 1st image. It maintains accurate segmentation even for small objects in the distance (the 3rd image). For fine structures such as hair, the competing methods often fail to achieve fine-level segmentation and tend to predict the surrounding areas as hair (the 2nd and 4th images. In contrast, MaSSFormer effectively distinguishes hair and other detailed elements from the background, ensuring high-quality segmentation.

D. Limitations

First, while MaSS13K provides 13K finely annotated images, it can be further expanded in the number of samples, scenes, and categories. Second, although MaSSFormer has achieved a relatively good balance between accuracy and efficiency, its computational cost and memory usage are still high, especially for mobile devices. New lightweight networks are expected for efficient yet accurate high-resolution semantic segmentation.

References

[1] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2

[2] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[3] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 2

[4] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. What is a good evaluation measure for semantic segmentation?. In *Bmvc*, pages 10–5244. Bristol, 2013. 2

[5] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[7] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[8] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3

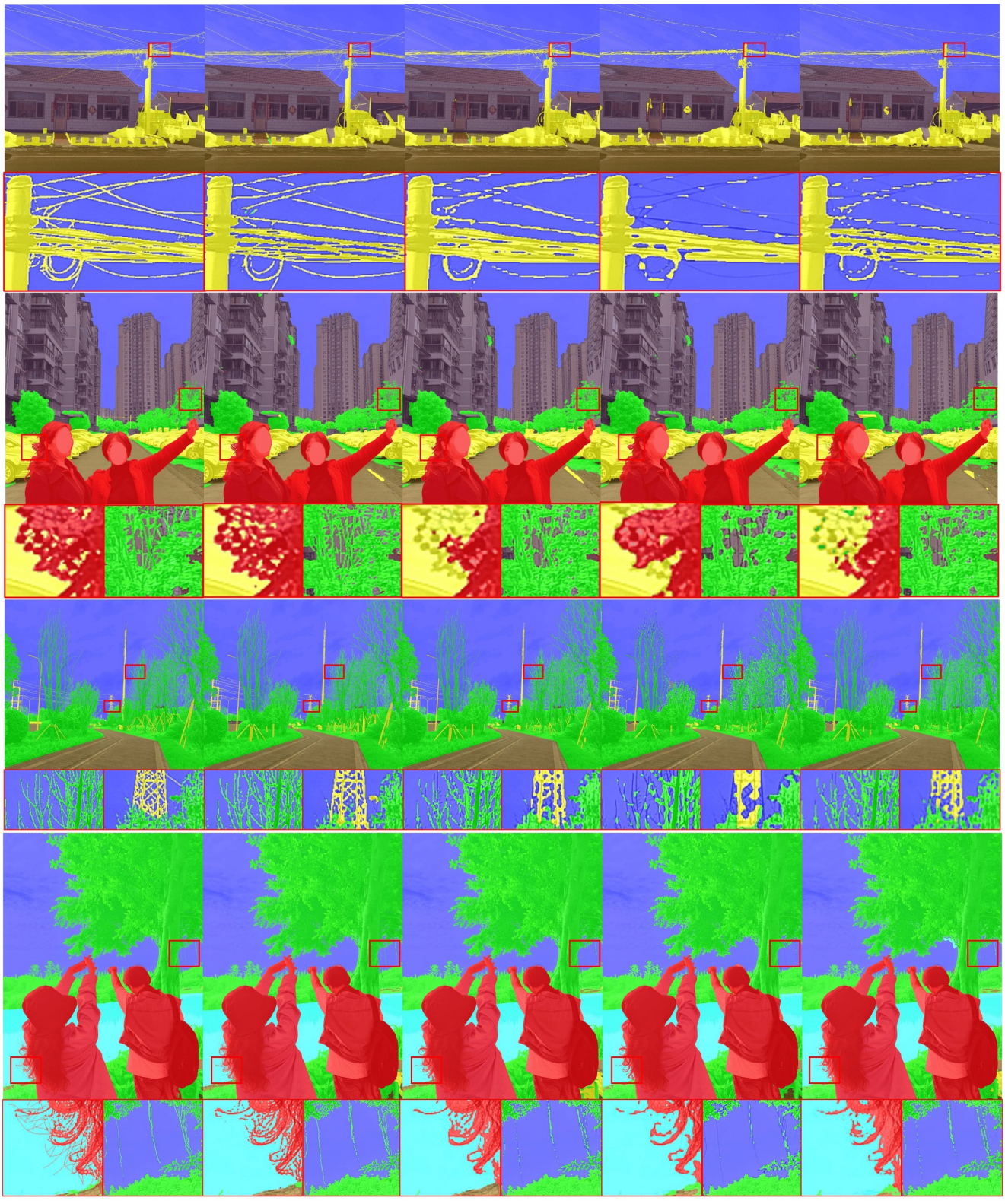

Figure 4. More qualitative comparisons between MaSSFormer and other methods. Please zoom-in for a better view.