MGMap: Mask-Guided Learning for Online Vectorized HD Map Construction

Supplementary Material

In this supplementary file, we provide more details, discussions, and experiments as follows:

- More details of our method;
- Additional experiments;
- Extensive qualitative results;
- Limitations and future work.

A. More Details of Our Method

A.1. Enhanced Multi-level Neck

In this section, we present more details of the fused attention in the enhanced multi-level (EML) neck. EML neck consists of three cascaded layers, each layer is a basic ResNet block [2] with fused channel attention (CA) and spatial attention (SA) [6], which focus on semantics and positional information to adaptively learn the crucial regions of the BEV space.

For the BEV feature with the shape of $D \times H_{BEV} \times W_{BEV}$, channel attention calculates channel weight of size $D \times 1 \times 1$ through average pooling and max pooling, emphasizing diverse channel information. A sigmoid function after MLP is employed to calculate the channel-wise attention map. For spatial attention, average pooling and max pooling are used to compress the channel dimension D, in which we construct spatial weight with shape $1 \times H_{BEV} \times W_{BEV}$ to obtain the weight of location information. Then, a spatial map can be generated by a 7×7 convolutional layer followed by the sigmoid function.

A.2. Auxiliary Loss Setting

In addition to regressing the point's position, an auxiliary loss for mask construction is required. As mentioned in the main paper, we combine cross-entropy loss $\mathcal{L}_{ce}$ and dice loss $\mathcal{L}_{dice}$ [4] to construct instance mask $\hat{\mathbf{M}}_{ins}$ and binary mask $\hat{\mathbf{M}}_b$. Specifically,

$$\mathcal{L}_{ins} = \lambda_{ins}\mathcal{L}_{ce}(\hat{\mathbf{M}}_{ins}, \mathbf{M}_{ins}) + \lambda_{d1}\mathcal{L}_{dice}(\hat{\mathbf{M}}_{ins}, \mathbf{M}_{ins}), \quad (1)$$

$$\mathcal{L}_{b} = \lambda_{b}\mathcal{L}_{ce}(\hat{\mathbf{M}}_{b}, \mathbf{M}_{b}) + \lambda_{d2}\mathcal{L}_{dice}(\hat{\mathbf{M}}_{b}, \mathbf{M}_{b}), \quad (2)$$

where $\{\lambda_{ins}, \lambda_{d1}, \lambda_b, \lambda_{d2}\}$ are the corresponding loss weights for two level masks. $\mathcal{L}_{ce}$ calculates the loss of each pixel equally, while $\mathcal{L}_{dice}$ takes consideration of mining the foreground areas, which can be formulated as below

$$\mathcal{L}_{dice} = 1 - 2 \cdot \frac{\hat{\mathbf{M}} \cap \mathbf{M}}{\hat{\mathbf{M}} \cup \mathbf{M}}. \quad (3)$$

To this end, we expect a larger intersection over the union area for the predicted mask $\hat{\mathbf{M}}$ and ground truth $\mathbf{M}$.

B. Additional Experiments

B.1. Time Analysis

Table A1 shows the detailed time analysis of each component. Compared with MapTR [3], EML neck and PG-MPR bring a slight time delay while the initial BEV extraction causes the main time-consuming.

Method	BEV Extractor init.BEV	BEV Extractor neck	decoder	PG-MPR	Total
MapTR	48.4ms	/	16.3ms	/	64.7ms
Ours	48.4ms	6.9ms	16.3ms	12.7ms	84.3ms

Table A1. Detailed runtime analysis and comparison with MapTR

B.2. Experiments under Different Conditions

We compare MGMap with the state-of-the-art method MapTR [3] under different weather and lighting conditions, in which the nuScenes dataset [1] is split by [5]. We employ ResNet-50 [2] as the image backbone and SECOND [7] for the LiDAR modality. All models are trained for 24 epochs. Moreover, experiments are conducted under mAP$_{chamfer}$ with a threshold setup of [0.5m, 1.0m, 1.5m]. As illustrated in Table A2, our method achieves the stable improvements with more than +9 mAP under different conditions.

Modality	Method	sunny	cloudy	rainy	day	night	mAP
Camera	MapTR	53.5	49.7	43.3	50.5	35.7	50.0
	MGMap	**65.2**	**62.8**	**49.4**	**61.7**	**37.6**	**60.8**
LiDAR	MapTR	59.2	56.9	46.7	56.3	38.9	55.8
	MGMap	**71.6**	**70.6**	**54.0**	**68.0**	**48.8**	**67.5**
Fusion	MapTR	66.6	62.7	54.5	63.4	44.8	62.8
	MGMap	**74.7**	**75.9**	**59.5**	**72.2**	**53.3**	**71.7**

Table A2. Comparisons under several weather and lighting conditions with different input modalities, our MGMap approach consistently achieves significant improvements over MapTR.

B.3. Ablations on Auxiliary Loss

In this section, we investigate the effects of auxiliary losses. As mentioned in the main paper, we present a parallel branch for mask predictions, which requires intensive supervision at the BEV space to construct the feature-prominent masks. Table A3 reports the experimental results. Compared with the model without auxiliary loss, mask construction introduces intensive pixel-level learning and alleviates the overfitting issue to some extent, resulting in a noteworthy +1.7 mAP improvement (57.6 *v.s.* 59.3). However, simple parallel segmentation learning lacks full use of mask features. To synergize with the vectorization task, our mask-guided design is proposed to boost the potential of mask features and obtains the best performance with 61.4 mAP.

Strategy	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
w/o. mask	53.1	60.0	59.5	57.6
+parallel segementation	53.3	63.0	61.6	59.3
+mask-guided design	**57.4**	**63.5**	**63.3**	**61.4**

Table A3. Ablation studies on auxiliary loss. Adding parallel segmentation achieves a certain level of improvement, while mask-guided design further enhances performance with the best result.

C. Extensive Qualitative Results

Figure A1 presents the visualization results of the learned masks and the final predictions, in which the binary masks are constructed to assist for the final map vectorization. Figure A2 provides the visual comparison with recent sotas. Later, Figures A3 to A6 provide extensive visualization results of our MGMap, comparing with the state-of-the-art approach MapTR under different weather and lighting conditions. Our MGMap method consistently demonstrates its promising capabilities across various scenarios.

D. Limitations and Future Work

As shown in Figure A6, under some adverse conditions, like low light, occlusion, and long-range perceptions, our image-based approach still has limitations in achieving reliable performance. It is mainly caused by the lack of effective features and the inferior interpretation of driving scenes. In the future, multi-modal fusion, temporal information, and the introduction of road priors will be explored to address the current shortcomings and obtain the vectorized HD map with higher precision.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[3] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. In *ICLR*, 2022. 1

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571. Ieee, 2016. 1

[5] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve. In *CVPR*, pages 13218–13228, 2023. 1

[6] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 1

[7] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1

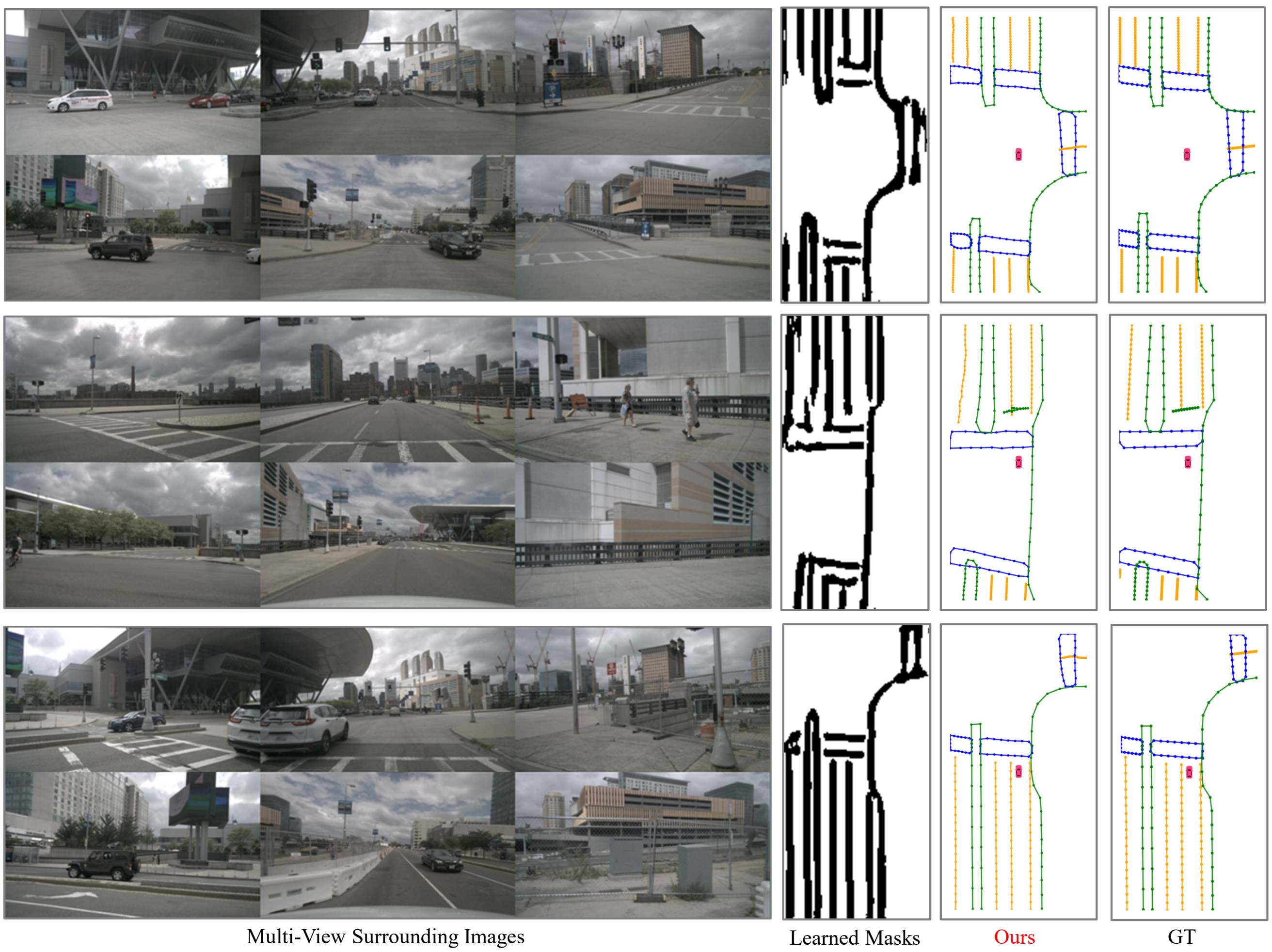


Figure A1. The visual results of the learned masks, our proposed **MGMap** approach and the corresponding ground truth.

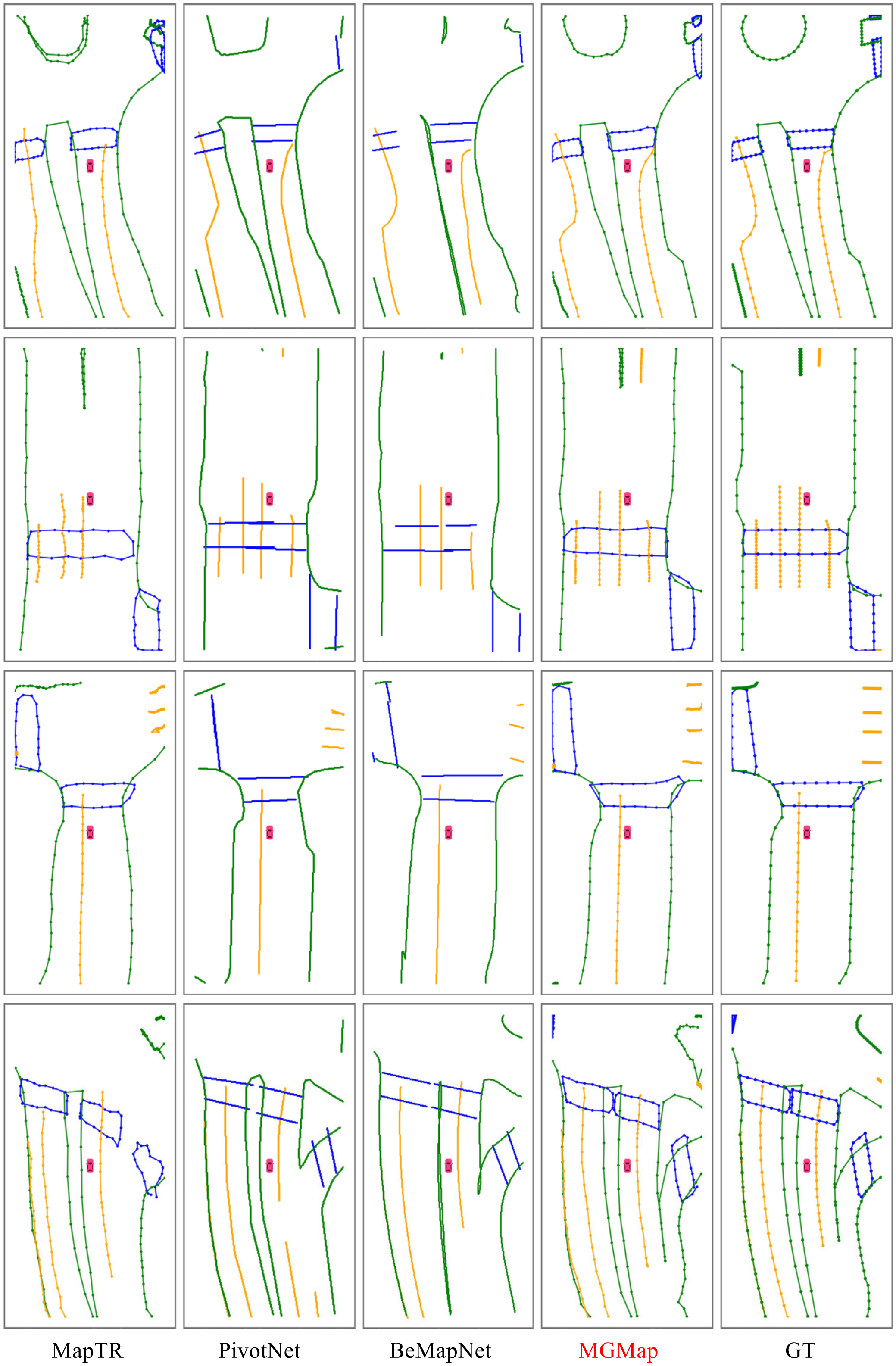


Figure A2. Comparison with recent sota methods.

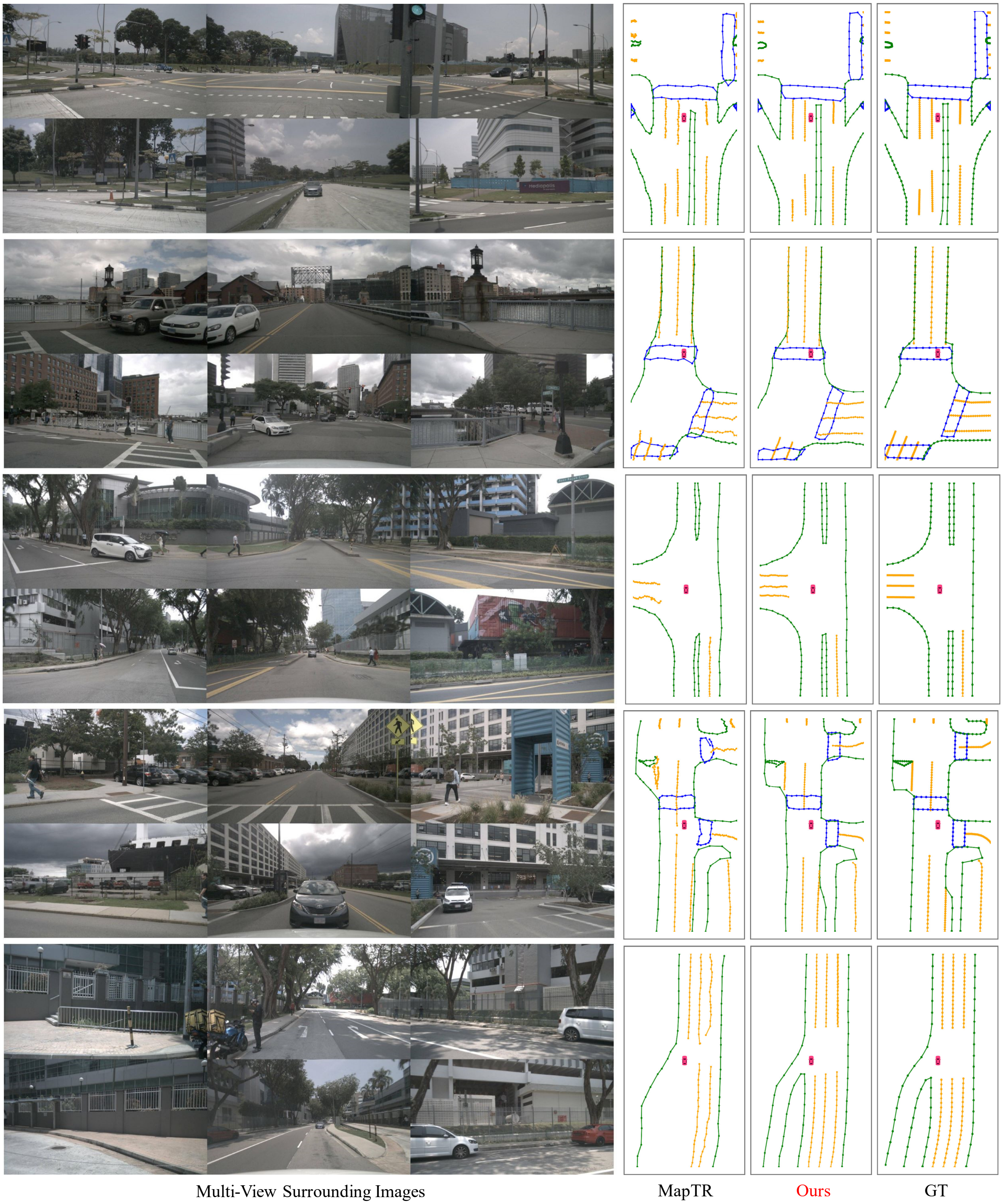
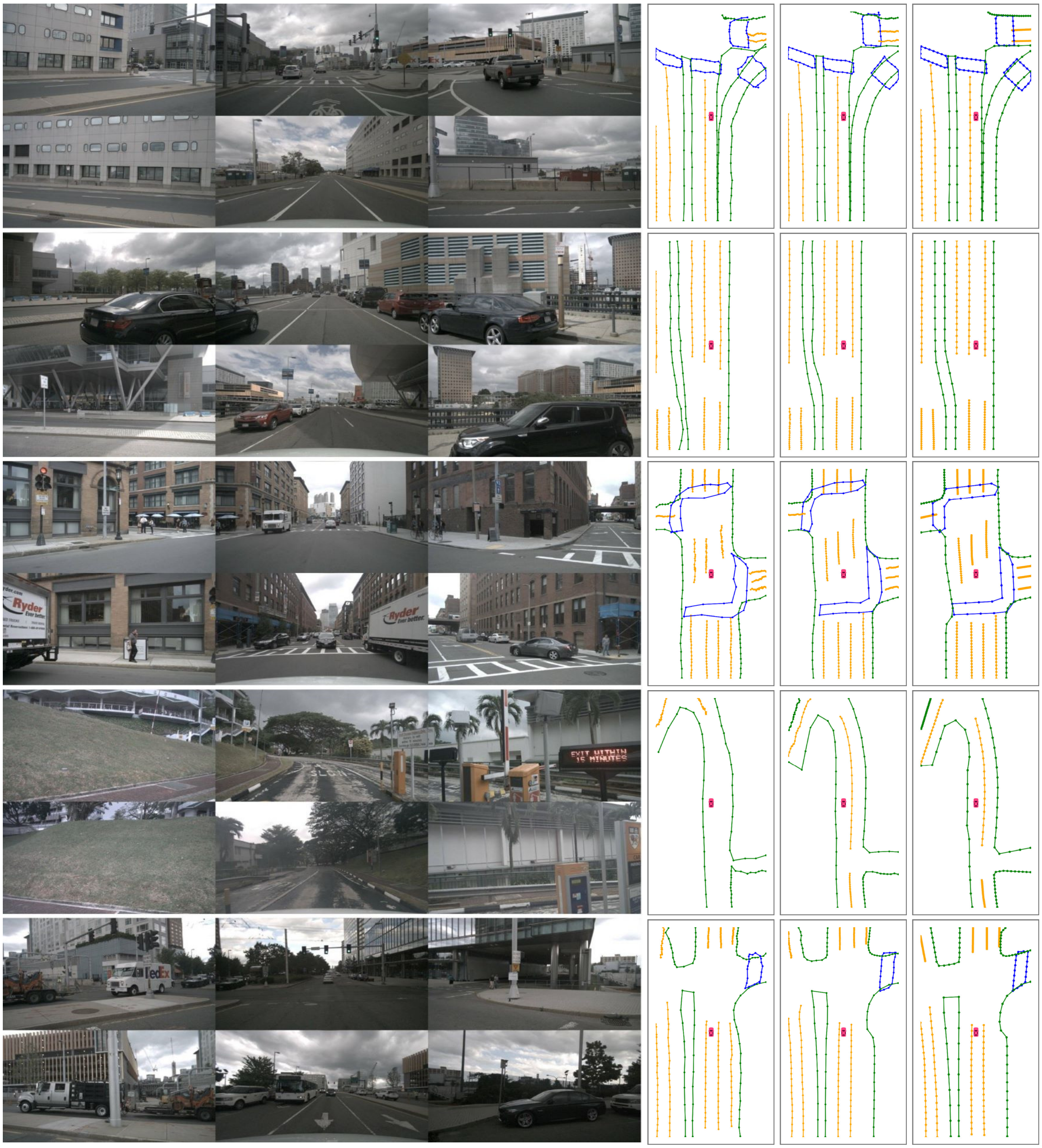


Figure A3. Visualization results under the weather condition of *sunny*.



Multi-View Surrounding Images MapTR Ours GT

Figure A4. Visualization results under the weather condition of *cloudy*.

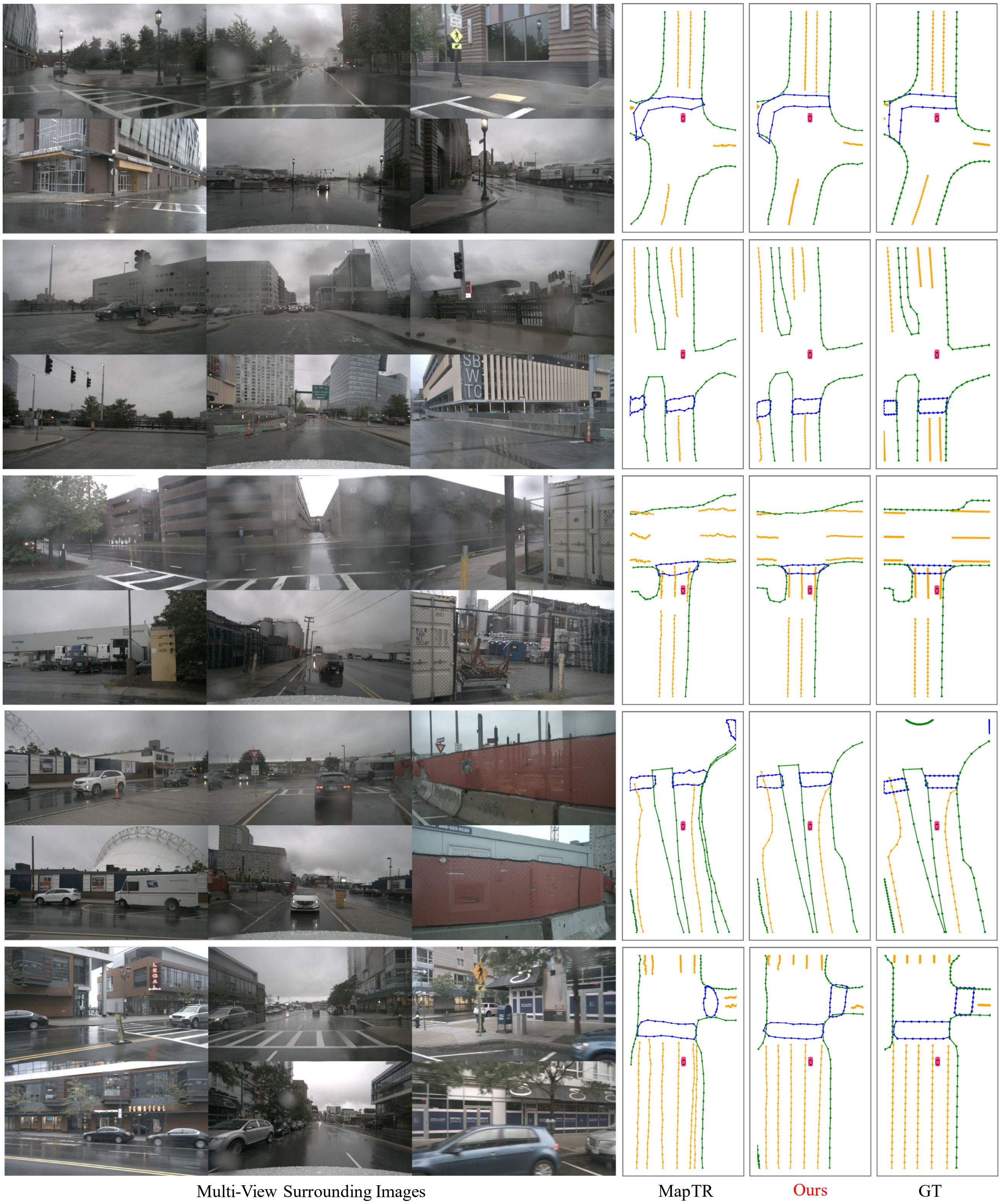


Figure A5. Visualization results under the weather condition of *rainy*.

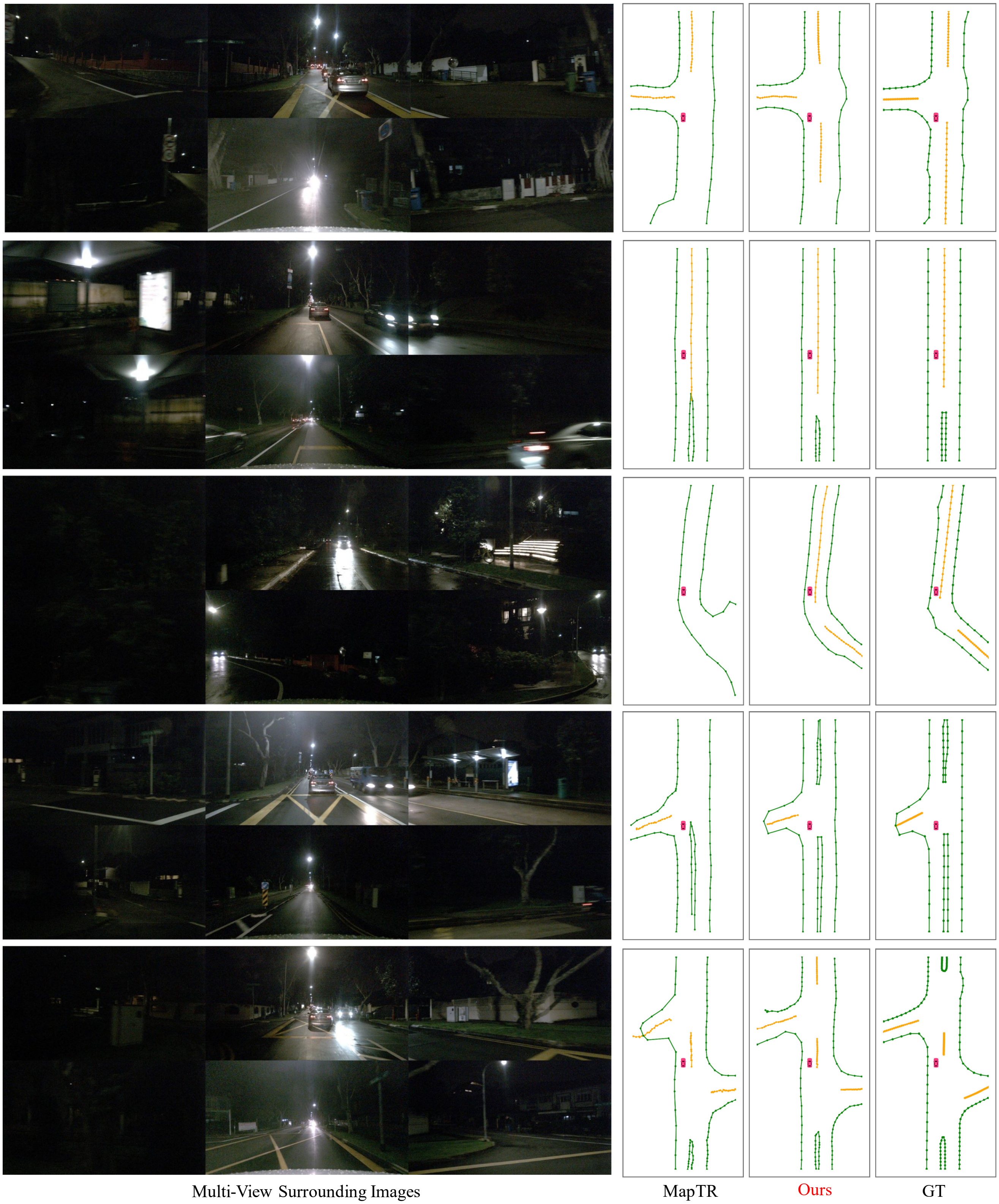


Figure A6. Visualization results under the *night* condition.