HIMap: HybrId Representation Learning for End-to-end Vectorized HD Map Construction Supplementary Material

Yi Zhou[1], Hui Zhang[1], Jiaqian Yu[1], Yifan Yang[1], Sangil Jung[2], Seung-In Park[2], ByungIn Yoo[2],
[1]Samsung R&D Institute China-Beijing (SRC-B)
[2]Samsung Advanced Institute of Technology (SAIT), South Korea
{yi0813.zhou, hui123.zhang, jiaqian.yu, yifan.yang}@samsung.com

In this supplementary material, we provide additional analysis of the proposed HIMap, including:

- More implementation details.
- Inference speed, memory, and model size.
- Extension: 3D Map and Centerline.
- More ablation studies on the backbone, the number of layers of the hybrid decoder, the number of points of an element, and the number of map elements.
- Additional examples of attention maps of HIQuery.
- Qualitative analysis on both nuScenes [1] and Argoverse2 [9] datasets.

A. Implementation Details

BEV Feature Extraction for Multi-modality Data. Given multi-modality inputs, *i.e.* multi-view RGB images and LiDAR point cloud data, we utilize a camera BEV feature extractor, a LiDAR BEV feature extractor, and a BEV feature fuser to generate BEV features. The camera BEV feature extractor is described in **??** of the main paper. LiDAR BEV feature extractor consists of a SECOND [10] backbone to extract sparse LiDAR features and a LiDAR-to-BEV projection module to generate LiDAR BEV features by flattening the sparse LiDAR features along the height dimension. Then the BEV feature fuser [3] concatenates the camera and LiDAR BEV features and utilizes a convolution layer to fuse them.

Prediction Heads. The class head, point head, and mask head consist of an FFN (two Linear layers) and an extra functional layer. The class head and point head utilize another Linear layer to predict the class and point coordinates respectively. The mask head generates the element mask by applying matrix multiplication with the element query inside HIQuery and the BEV features.

Training. We utilize the BEVFormer [2] encoder as the 2D-to-BEV feature transformation module and set the size of each BEV grid to $0.3m$ by default. The default number for map elements, points in an element, and layers of the hybrid decoder is $50, 20, 6$, respectively. For all experiments, distributed training with 8 GPUs is utilized and the total batch size is 32. The optimizer, learning rate scheduler, base learning rate, and weight decay are set to AdamW [7], Cosine Annealing, $0.0006, 0.01$, respectively. We employ the Hungarian matching algorithm as matching criteria to obtain the unique assignment between predictions and GTs. The matching cost used by Hungarian matching integrates the matching losses of class probabilities, point coordinates, point directions, and masks. For loss supervision of prediction heads, the class head is supervised with focal loss. The point head is with point position (L1 loss) and direction (Cosine Embedding loss) losses. The mask head is with binary cross-entropy loss and dice loss.

	MapTR [3]	MapTRv2 [4]	BeMapNet [8]	HIMap (ours)
FPS	21.6	18.7	9.7	11.4
GPU mem.(MB)	2544	2888	5484	3512
Params (MB)	35.9	40.3	73.8	63.2
mAP	59.3 (-14.4)	68.7 (-5.0)	64.8 (-8.9)	**73.7**

Table S1. Comparison with SOTA methods on nuScenes val set. FPSs are measured on one A100 GPU with batch size as 1.

Inference. Given multi-view RGB images or multi-modality inputs, HIMap directly predicts class, point coordinates, and masks of map elements. The first two kinds of outputs are utilized for calculating the mAP result. Masks are optional for producing rasterized HD map. Without any post-processing, the top-scoring predictions are taken as final results.

B. Inference Speed, Memory, and Model Size.

Comparison with SOTA methods in the above aspects are shown in Table S1. (1) Compared with BeMapNet [8], HIMap achieves 8.9 mAP gain with *faster* speed, *fewer* parameters, and *smaller* GPU memory cost. (2) Compared with MapTRv2 [4], HIMap obtains 5.0 mAP gain with lower efficiency. As we discussed in the Limitation part, this paper mainly focuses on improving the map reconstruction accuracy. We believe that HIMap boosts the per-

Methods	Epoch	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
		easy: $\{0.5, 1.0, 1.5\}$m			
VectorMapNet[5]	24	36.5	35.0	36.2	35.8
MapTRv2 [4]	6	60.7	**68.9**	64.5	64.7
Ours	6	**66.7**	68.3	**70.3**	**68.4** (+3.7)

Table S2. **Comparison to the state-of-the-art on Argoverse2 *val* set with 3D map predictions.**

Methods	Epoch	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	$AP_{cen.}$	mAP
		easy: $\{0.5, 1.0, 1.5\}$m				
MapTRv2 [4]	6	53.5	**66.9**	63.6	61.5	61.4
Ours	6	**64.6**	66.4	**71.1**	**66.6**	**67.2** (+5.8)

Table S3. **Comparison to the state-of-the-art on Argoverse2 *val* set with 3D map predictions and centerline learning.**

Modality	Backbone	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
		easy: $\{0.5, 1.0, 1.5\}$m			
C	ResNet50	71.3	75.0	74.7	73.7
	Swin-Tiny	72.3	75.9	76.3	74.8
C + L	ResNet50 & Second	77.0	74.4	82.1	77.8
	Swin-Tiny & Second	78.7	75.7	83.3	79.3

Table S4. **Ablations about Swin [6] backbone on nuScenes *val* set.** "C" and "L" refers to Camera and LiDAR respectively.

layer number	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
1	57.2	65.6	63.6	62.2
2	67.4	71.2	71.3	70.0
3	69.5	72.1	73.1	71.6
6	**71.3**	**75.0**	**74.7**	**73.7**
8	69.7	71.9	73.9	71.9

Table S5. **Influence of layer number of hybrid decoder.**

formance to an unprecedented level. Such kind of high-accuracy models have essential values for many application scenarios, *e.g.* offline HD map construction, auto labeling system *etc*. Some techniques, *e.g.* model quantization, pruning, and distillation, could be explored to improve the efficiency in future work.

C. Extension: 3D Map and Centerline.

Since Argoverse2 dataset [9] provides 3D vectorized map annotations, we further extend HIMap to the 3D map construction. A set of learnable 3D anchor points are utilized and 3D point coordinates are directly predicted by the point head. As shown in Table S2, on the 3D HD map construction task, HIMap also consistently exceeds previous SOTAs. What's more, we further predict more categories of elements, *e.g.* centerline, in the 3D map. As shown in Table S3, with centerline included, HImap outperforms MapTRv2 [4] by 5.8 mAP.

D. More Ablation Study

Swin Transformer Backbone. We study the effect of utilizing Swin-Tiny [6] backbone with different input modality and show the results in Table S4. As we can see, replacing ResNet50 with the Swin-Tiny backbone can further improve the performance of HIMap. With both camera images and LiDAR point cloud data, HIMap achieves 79.3 mAP.

point number	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
5	48.5	72.6	60.1	60.4
10	68.8	74.1	73.1	72.0
20	**71.3**	**75.0**	**74.7**	**73.7**
30	72.1	73.8	75.0	73.7
40	70.1	70.0	73.9	71.3

Table S6. **Influence of point number.** The element number is set to 50.

element number	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
35	69.5	72.9	72.7	71.7
50	**71.3**	**75.0**	**74.7**	**73.7**
75	70.5	71.8	73.8	72.1
100	70.6	72.8	74.3	72.6

Table S7. **Influence of element number.** The point number is set to 20.

Layer Number of Hybrid Decoder. We present the results of different number of layers of the hybrid decoder in Table S5. The results continue to improve as the number of layers increases and reach saturation when utilizing six layers.

Number of Points. The influence of different point number of an element (*i.e.* P) is shown in Table S6. Empirically, we find utilizing 20 points achieves the best performance. We speculate that too few points are insufficient to express the details of the element, while too many points increase the optimization difficulty and reduce accuracy.

Number of Elements. The influence of different number of elements (*i.e.* E) is shown in Table S7. Too small element number intensifies the competition between elements for HIQuery, while too large element number introduces more False-Positive (FP) and drops the performance. We set the element number to 50 empirically.

E. More Attention Maps of HIQuery

In Figure S1, we provide more attention maps of anchor points with its sampling points and anchor masks for a single map element. These visualizations validate that anchor points and masks focus on local and overall information of elements respectively, and point-element interaction helps to achieve mutual refinement.

F. Qualitative Analysis

In Figure S2 and S3, we show the result comparison between BeMapNet [8], MapTRv2 [4], and the proposed HIMap on the nuScenes [1] dataset. In Figure S4, we present the result comparison between MapTRv2 [4] and the proposed HIMap on the Argoverse2 [9] dataset. Our HIMap generates impressive results in various driving scenes. Compared with BeMapNet [8] and MapTRv2 [4],

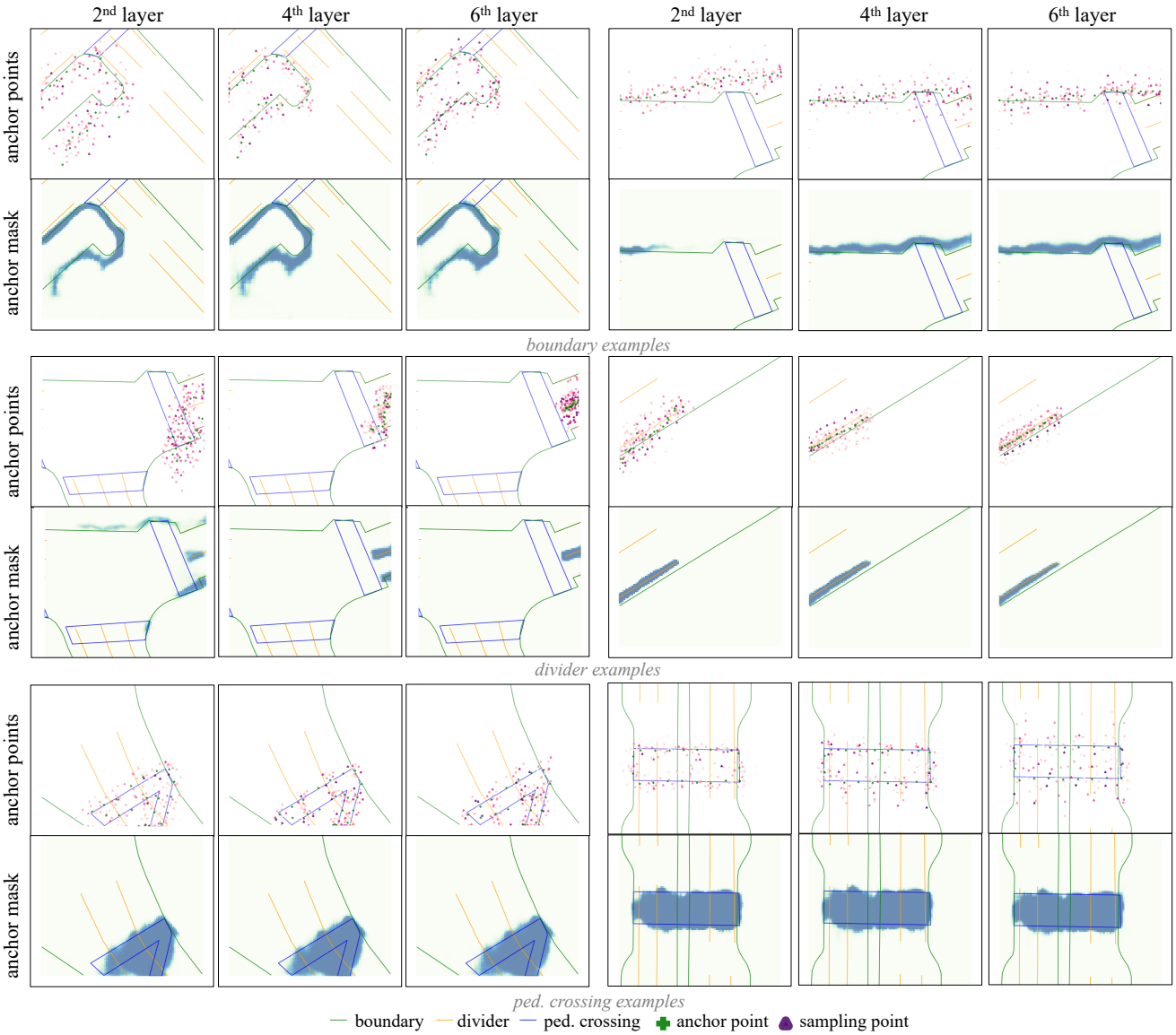

Figure S1. **Attention maps of HIQuery at different layers.** Attention maps are overlaid on the GT. The darker the color, the greater the attention value. Best zoom-in and viewed in color.

our results have richer details, more accurate shape and point positions of map elements, and avoid inter-element entanglement.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[2] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 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. *arXiv preprint arXiv:2208.14437*, 2022. 1

[4] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736*, 2023. 1, 2

[5] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*,

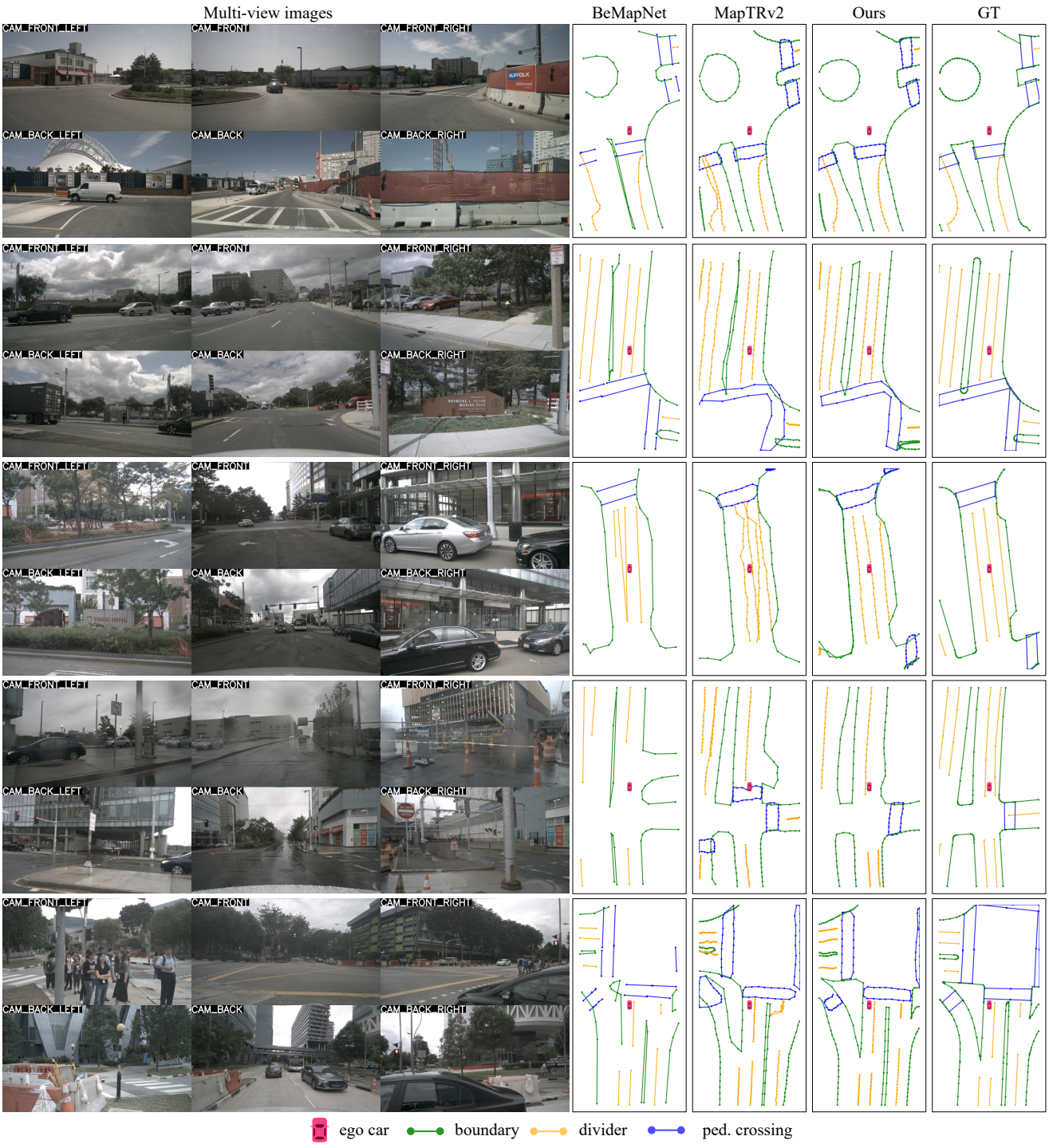

Figure S2. **Qualitative result comparison on nuScenes dataset.** From left to right: input multi-view images, BeMapNet predictions, MapTRv2 predictions, our predictions, and GT annotation. Each row corresponds to one sample. For BeMapNet predictions, the semi-closed or closed boundaries easily have shrunk shapes (1st, 2nd, 4th, 5th samples), the length of the divider is inaccurate in 3rd and 4th samples, and the ped crossing is missing or has an incomplete shape in 3rd, 4th, 5th samples. For MapTRv2 predictions, the shape of boundary is inaccurate in 2nd, 3rd, 4th, 5th samples, dividers are entangled in 1st, 3rd samples, and the ped crossing is missing in 3rd, 5th samples. In comparison, our results have more accurate point positions and shapes of map elements, and avoid inter-element entanglement. Best viewed in color.

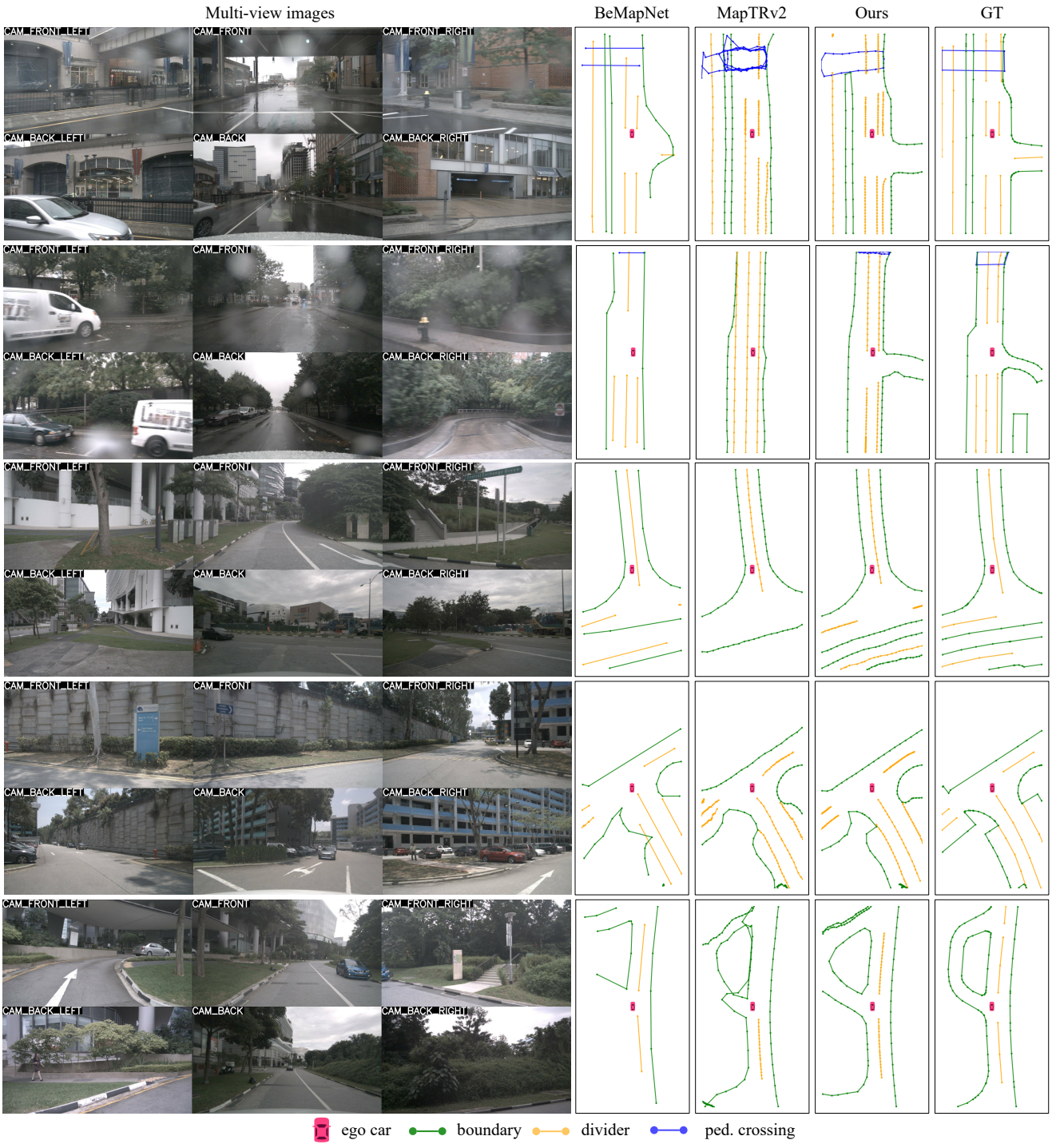

Figure S3. **Qualitative result comparison on nuScenes dataset.** From left to right: input multi-view images, BeMapNet predictions, MapTRv2 predictions, our predictions, and GT annotation. Each row corresponds to one sample. For BeMapNet predictions, the shape of the boundary is inaccurate in 1st, 2nd, 4th, 5th samples, the length of divider is inaccurate in 1st, and 3rd samples. For MapTRv2 predictions, the shape of the boundary is inaccurate in 1st, 2nd, 4th, 5th samples, the length of the divider is inaccurate in 1st, and 2nd samples, and the divider and boundary are missing in 3rd samples. In comparison, our results have richer details, and more accurate point positions and shapes of map elements. Best viewed in color.

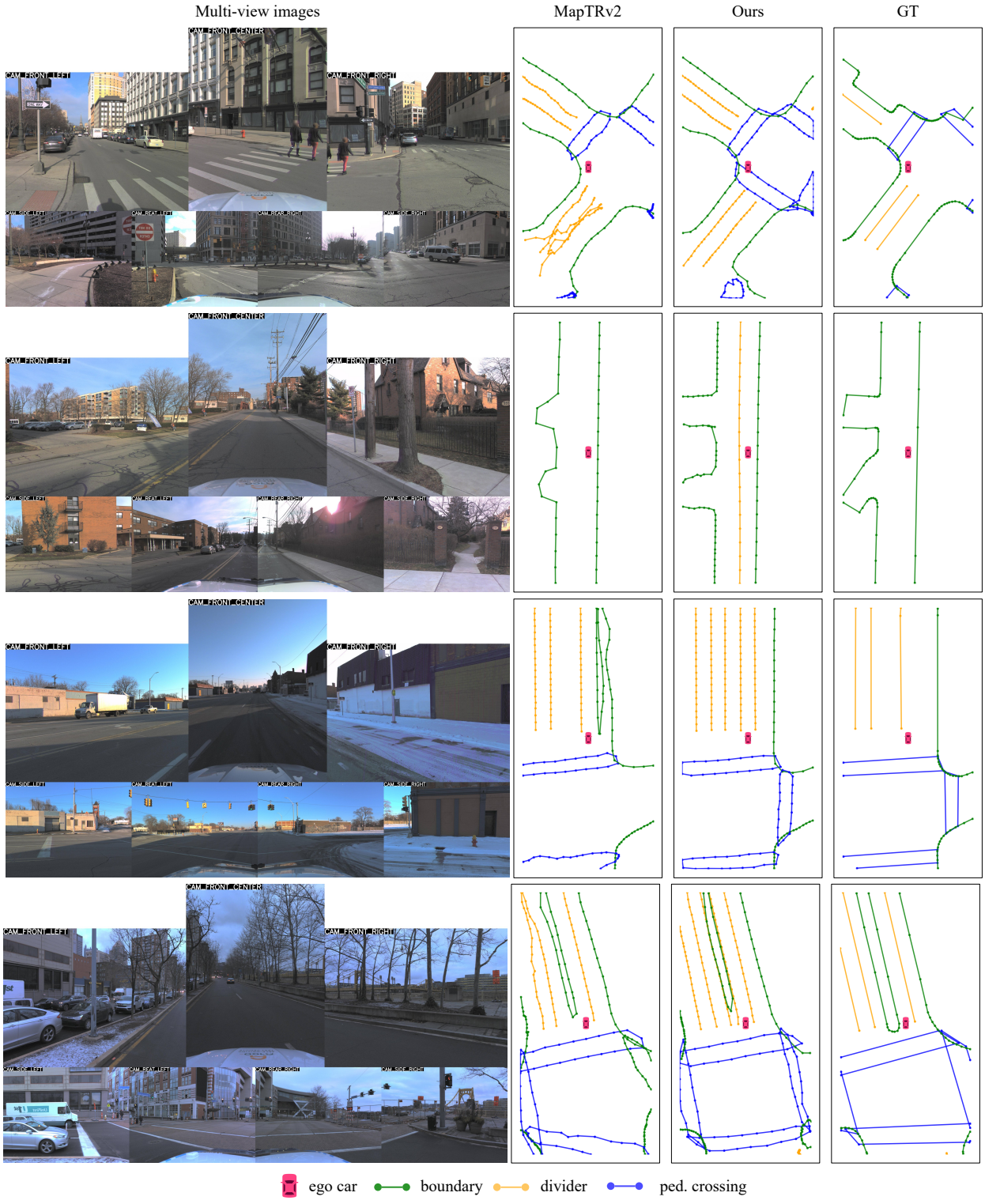

Figure S4. **Qualitative result comparison on Argoverse2 dataset.** From left to right: input multi-view images, MapTRv2 predictions, our predictions, and GT annotation. Each row corresponds to one sample. For MapTRv2 predictions, the shape of the boundary is inaccurate in 2nd and 4th samples, dividers are entangled in 1st and 4th samples, and ped crossing is missing in 3rd and 4th samples. In comparison, our results have more accurate point positions and shapes of map elements, and avoid inter-element entanglement. Best viewed in color.

pages 22352–22369. PMLR, 2023. 2

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[8] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13218–13228, 2023. 1, 2

[9] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 1, 2

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