Supplementary Material for Enhancing Vectorized Map Perception with Historical Rasterized Maps

Xiaoyu Zhang[1,*], Guangwei Liu[2,*], Zihao Liu[3], Ningyi Xu[3], Yunhui Liu[1,✉], and Ji Zhao[2,†]

[1] The Chinese University of Hong Kong
[2] Huixi Technology
[3] Shanghai Jiao Tong University
zhang.xy@link.cuhk.edu.hk xuningyi@sjtu.edu.cn
yhliu@mae.cuhk.edu.hk zhaoji84@gmail.com

Appendix

A Performance under Original MapTRv2 Setting

Here we provide the validation results on Argoverse 2 dataset in Tab. 1, under the original MapTRv2 setting. Here all training data at 10 Hz is used for training, and 2.5 Hz validation data is extracted for validation. Our HRMapNet still outperforms the baseline, MapTRv2.

Table 1: Comparison on Argoverse 2 [6] under the original MapTRv2 setting.

Method	Epoch	AP_{ped}	AP_{div}	AP_{bou}	mAP	FPS
MapTRv2 [5]	6	62.9	**72.1**	67.1	67.4	18.1
# HRMapNet	6	**64.1**	71.5	**69.7**	**68.5**$^{(\uparrow 1.1)}$	16.2

B Performance under Challenging Conditions

To further demonstrate the improvements of utilizing a HRMap, we summarize the performance under different conditions in Tab. 2. It is clear the improvement is significant especially for these challenging conditions.

*: Equal contribution. †: Project lead. ✉: Corresponding author.

Table 2: Performance under different conditions on nuScenes [1], trained with 24 epochs.

Method	Initial Map	Night	Rainy	Normal
MapTRv2	-	39.6	50.8	64.6
# HRMapNet	Empty	42.6$^{(\uparrow 3.0)}$	62.6$^{(\uparrow 11.8)}$	69.4$^{(\uparrow 5.2)}$
	Training Map	74.8$^{(\uparrow 35.2)}$	72.9$^{(\uparrow 22.1)}$	85.9$^{(\uparrow 21.3)}$

C Performance for Long Range

We test HRMapNet (StreamMapNet [7] as baseline) by just increasing perception range. As in Tab. 3, with online constructed map, HRMapNet also boosts the baseline a lot. For such far ranges, we suggest loading a pre-built map like in P-MapNet [2] for practical usage. We further test HRMapNet with a pre-built map from training data, and mAP is improved to 57.3 and 40.2, respectively. Besides, we believe HRMapNet can achieve better results with more careful settings, such as suitable map resolution, query number.

Table 3: Evaluation under long range on nuScenes [1]. The results of MapTR and P-MapNet are from the paper of P-MapNet. StreamMapNet is trained and evaluated using the official codes. Our HRMapNet is based on StreamMapNet. †: Performance with a loaded map built from training data. All models are trained with 24 epochs.

Range	Method	AP_{ped}	AP_{div}	AP_{bou}	mAP
120×60m	MapTR [4]	18.9	26.0	15.7	20.2
	# P-MapNet [2]	22.0	27.2	19.5	22.9
	StreamMapNet [7]	37.2	42.3	30.2	36.6
	# HRMapNet	**43.1**	**47.7**	**34.9**	**41.9**$^{(\uparrow 5.3)}$
	# HRMapNet†	65.0	61.9	44.9	57.3
240×120m	MapTR [4]	7.2	12.7	4.2	8.0
	# P-MapNet [2]	16.3	22.7	10.5	16.5
	StreamMapNet [7]	22.4	31.3	16.0	23.3
	# HRMapNet	**29.4**	**34.2**	**19.7**	**27.8**$^{(\uparrow 4.5)}$
	# HRMapNet†	51.0	46.2	23.5	40.2

D Ablation on Decoder Layers

Our method keeps the map decoder module the same as the baseline methods [5, 7], and 6 decoder layers are used in these methods. With the help of retrieved rasterized map as priors, the proposed query initialization module helps map element queries search for desirable map elements more efficiently. Here, we

provide an ablation study of the number of decoder layers in Tab. 4 to further demonstrate the searching efficiency improved by this module. The results of not using query initialization are also listed for comparison.

Increasing decoder layers commonly brings higher accuracy. The comparison indicates using only 4 decoder layers with query initialization have already achieved good results, better than the method using 6 decoder layers without query initialization. It is because query initialization endows map element queries with priors.

Table 4: Ablation of the number of decoder layers on nuScenes [1], trained with 24 epochs. The results of the method without query initialization are listed for comparison. The results better than the method without query initialization are highlighted in blue.

Method	Layer	AP_{ped}	AP_{div}	AP_{bou}	mAP
w/o Query Initialization	6	62.6	64.5	66.8	64.6
w/ Query Initialization	3	60.7	61.9	65.4	62.6
	4	63.6	64.7	68.0	65.4
	5	64.5	66.5	68.6	66.6
	6	65.8	67.4	68.5	67.2

E Discussion about Localization Error

By default, we train and test HRMapNet with the groundtruth ego-pose, which is a common practice in temporal-based map perception and object detection methods, such as StreamMapNet [7] and BEVFormer [3]. In the main body of the paper, it is demonstrated that HRMapNet has certain robustness to pose error even without specific design. The noise level is set according to common requirements in self-driving.

To further deal with localization error in practice, we can add pose noise as augmentation during training, and the results are listed in Tab. 5. The robustness is further enhanced even for large pose errors, with some compromise of accuracy. Furthermore, as a future work, we can change convolution-based to attention-based BEV feature aggregation to alleviate misalignment caused by large localization errors.

Table 5: Performance with the augmentation of adding pose noise on nuScenes [1], trained with 24 epochs. σ_r is applied standard deviations of rotation noise.

σ_r (rad)	0	0.005	0.01	0.02	0.05	0.1
HRMapNet	65.8	65.7	65.4	65.4	64.8	63.1

F Visualized Prediction Results

In Fig. 1, Fig. 2, and Fig. 3, we provide more visualized comparison results. All three methods are trained with 24 epochs. Ours is based on MapTRv2. The last column are visualized retrieved local rasterized maps. With the help of the local map, HRMapNet commonly achieves more accurate results, such as predicting map elements which are hard to recognize in images because of occlusion, bad weather or at long range.

We evaluate constructed rasterized map using mean intersection over union (mIOU) and the result is mIOU=35.6 (ped:48.9, div:26.6, bou:31.5), which is not a high performance. But our target is not to build and utilize a perfect HRMap. Although occasionally retrieved local maps are with noise or even error, HRMapNet still predicts good results from images. Such examples are mainly illustrated in Fig. 3. The map noise or error is mainly from previous noisy predictions because of adverse weather, bad illumination or occlusions at turning or long range. Since such noisy predictions take only a small part, these map noise would be removed gradually by multiple accurate predictions in long-term running. Specifically, for the example in the first row of Fig. 3, the global map is still empty since it is the first frame, but HRMapNet still provides accurate results. For the example in the second row of Fig. 3, the prediction results are not affected by the errors in the retrieved local map, caused by previous predictions. And such map error is gradually removed by new accurate predictions, as illustrated in the third row.

G Visualized Global Rasterized Map

We provide some visualized examples of the final global rasterized maps in Fig. 4 and Fig. 5. These global maps are constructed from empty ones, and updated gradually by prediction results of testing data. We also provide a supplementary video for this process.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
2. Jiang, Z., Zhu, Z., Li, P., Gao, H.a., Yuan, T., Shi, Y., Zhao, H., Zhao, H.: P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. arXiv preprint arXiv:2403.10521 (2024)
3. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European Conference on Computer Vision (ECCV). pp. 1–18 (2022)
4. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: MapTR: Structured modeling and learning for online vectorized HD map construction. In: International Conference on Learning Representations (ICLR) (2023)

5. Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., Wang, X.: Maptrv2: An end-to-end framework for online vectorized hd map construction. arXiv preprint arXiv:2308.05736 (2023)
6. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2021)
7. Yuan, T., Liu, Y., Wang, Y., Wang, Y., Zhao, H.: Streammapnet: Streaming mapping network for vectorized online hd map construction. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7356–7365 (2024)

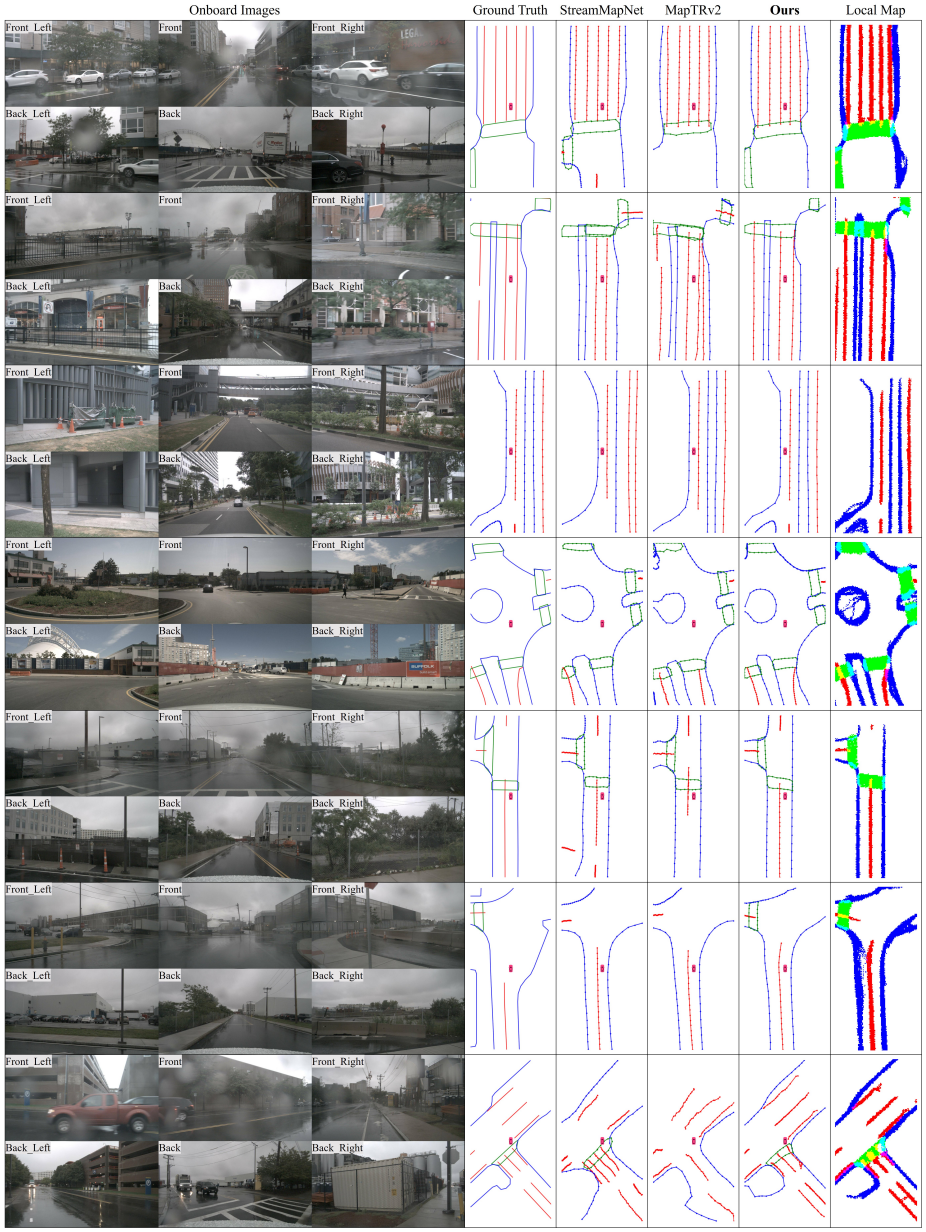

Fig. 1: Visualized results. The last column is the retrieved local rasterized maps in HRMapNet. Lane divider, pedestrian crossing and road boundary are illustrated in red, green and blue respectively.

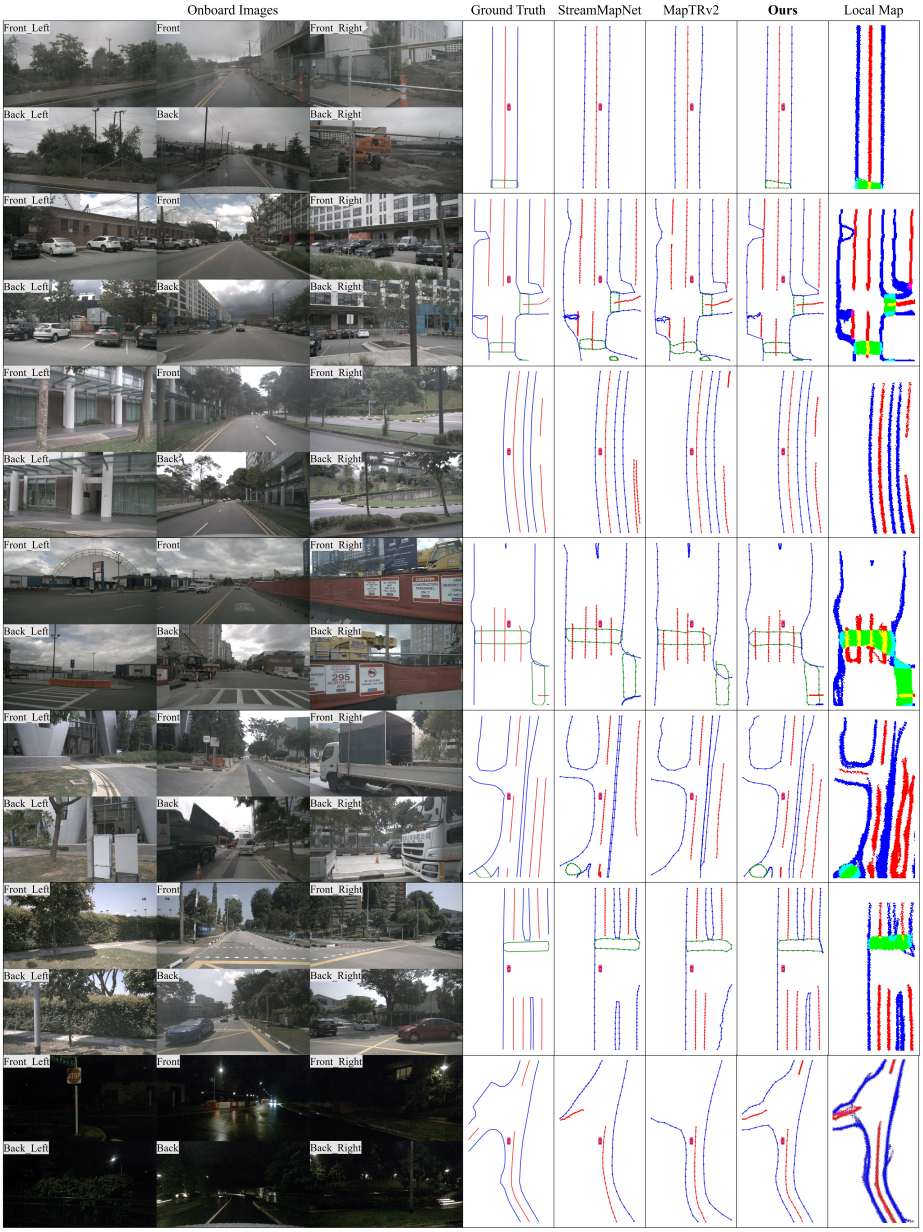

Fig. 2: Visualized results. The last column is the retrieved local rasterized maps in HRMapNet. Lane divider, pedestrian crossing and road boundary are illustrated in red, green and blue respectively.

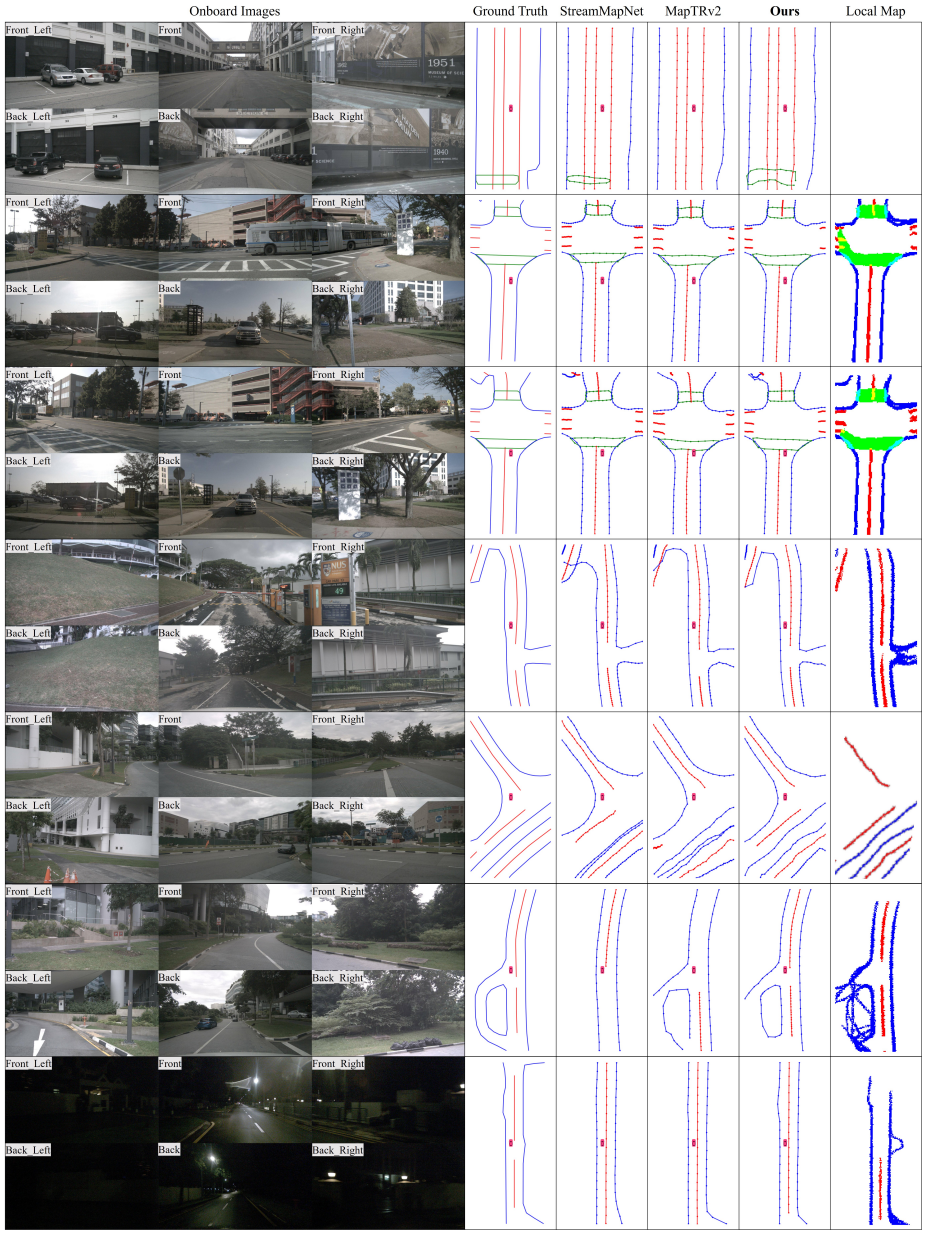

Fig. 3: Visualized results. The last column is the retrieved local rasterized maps in HRMapNet. Lane divider, pedestrian crossing and road boundary are illustrated in red, green and blue respectively.

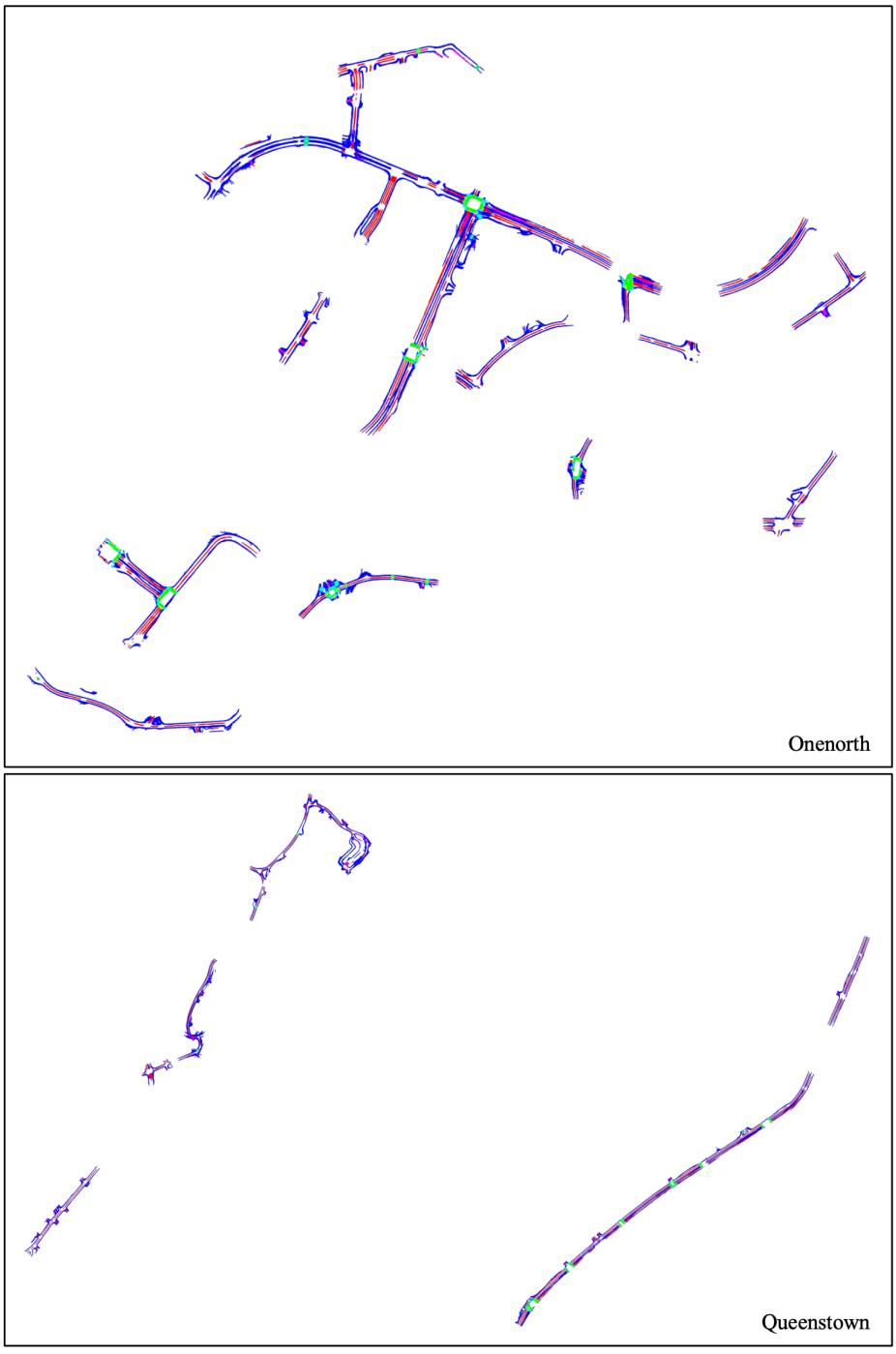

Fig. 4: Visualized global rasterized maps.

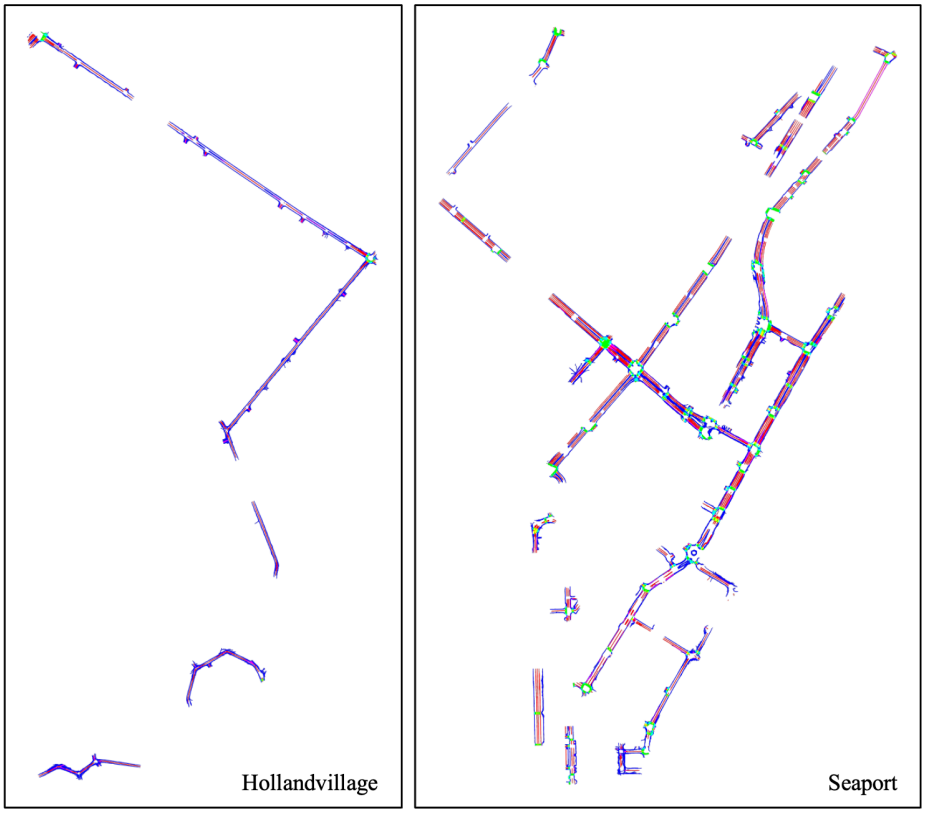

Fig. 5: Visualized global rasterized maps.