Uncertainty-Instructed Structure Injection for Generalizable HD Map Construction

Supplementary Material

In this supplementary material, we provide more details on our implementation and experiments as follows.

- Section A: More details on dataset partitions;
- Section B: More details on implementation;
- Section C: Additional experiments;
- Section D: Qualitative results and failure case analysis;
- Section E: Limitations and future work.

A. Dataset Partitions

To evaluate the generalization capability for online HD map vectorization, we conducted experiments on the geospatial disjoint (geo-based) dataset partitions. Here are some statistical analyses of the datasets. Table A1 shows the number of samples and overlap ratios for the original and region-based data partitions. Compared to widely-used original splits, region-based partition divides datasets according to the geographic location with lower overlap rates, with 11% for nuScenes [1] and 0% overlap for Argoverse2 [5], which are measured for all locations within a 30m radius around the ego-vehicles. For the city-based partition following [3], Table A2 shows the city-based distribution of the training and validation data. All dataset partitions are maintained on roughly the same scale as the original ones, in which Argoverse2 is resampled by 1/5 (2Hz) to be consistent with nuScenes dataset.

	Split	Train #	Val. #	Overlap Ratio
nuScenes	Ori.	27968	6019	85%
	Region	27846	5981	**11%**
Argoverse2	Ori.	21794	4704	54%
	Region	23434	4676	**0%**

Table A1. Data collections on region-based partitions.

	nuScenes		Argoverse2	
	City	Sample #	City	Sample #
Train	Boston, Onenorth	25926	Miami, Pittsburgh	21975
Val	Queentown, Holland Village	8056	Austin, Detroit Washington	9232

Table A2. Data collections on city-based partitions.

B. More Implementation Details

This section introduces the detailed settings of pre-trained perspective-view (PV) detection branches, including the training process and the transformation of the ground truth of PV-level detection.

B.1. Settings of PV Detection Branch

As described in the main paper, we introduce the pre-trained PV detection branch for explicit structural priors. In the design of the PV branch, given the total number of I images captured by the onboard cameras, we utilize the shared ResNet50 [2] backbone followed by the FPN [4] neck to extract multi-scale image features. Then, a multi-layer uncertainty-aware (UA) decoder is employed to detect map elements, in which we replace the original Deformable-DETR [6] with the UA-decoder design to take consideration of reliable PV coordinate output and uncertainty information, which will be further deployed in our UI2DPrompt module. We trained the model on four A100 GPUs with a batch size of 4, which refers to $4 \times I$ images in one batch. The learning rate is set to $3e^{-4}$ and the PV models are trained for 24 epochs. The loss setting is similar to the main branch of the BEV map, which contains $\mathcal{L}^{\text{pv}}_{\text{nll}}$ fused with the loss of the point regression $\mathcal{L}^{\text{pv}}_{\text{pts}}$ to produce uncertainty of the Laplace distribution and stabilize the coordinate output, in addition to $\mathcal{L}^{\text{pv}}_{\text{cls}}$ for classification. Thus, the PV loss can be formulated as below:

$$\mathcal{L}^{\text{pv}}_{\text{map}} = \lambda^{\text{pv}}_{\text{pts}}\mathcal{L}^{\text{pv}}_{\text{pts}} + \lambda^{\text{pv}}_{\text{nll}}\mathcal{L}^{\text{pv}}_{\text{nll}} + \lambda^{\text{pv}}_{\text{cls}}\mathcal{L}^{\text{pv}}_{\text{cls}}, \tag{1}$$

where the corresponding loss weights $\lambda^{\text{pv}}_{\text{nll}}$, $\lambda^{\text{pv}}_{\text{pts}}$, and $\lambda^{\text{pv}}_{\text{cls}}$ are set to 0.05, 50.0, and 5.0, respectively.

B.2. Ground Truth for PV Detection

Since the map annotations are labeled in the bird's-eye-view (BEV) space, the PV map labels are obtained through the projection of the BEV map ground truth. Given map ground truth at the ego-coordinate system (p_x, p_y), map polylines are transformed into the image-coordinate system with 2D coordinates $(x_{\text{pv}}, y_{\text{pv}})$ by camera extrinsic $\text{T}_{\text{ego2cam}}$ and intrinsic K_{cam}, which can be formulated as:

$$\mathbf{P}^{\text{h}}_{\text{cam}} = \text{T}_{\text{ego2cam}} \cdot \begin{bmatrix} p_x \\ p_y \\ 0 \\ 1 \end{bmatrix}, \tag{2}$$

$$\begin{bmatrix} x_{\text{pv}} \\ y_{\text{pv}} \\ 1 \end{bmatrix} = \frac{1}{z_{\text{cam}}} \cdot \mathbf{K}_{\text{cam}} \cdot \mathbf{P}^{\text{3D}}_{\text{cam}}, \tag{3}$$

where $\mathbf{P}^{\text{3D}}_{\text{cam}}$ is the first three dimension of $\mathbf{P}^{\text{h}}_{\text{cam}}$. The depth of the camera coordinates z_{cam} is used for normalization when projecting a point from the 3D space (camera coordinate system) onto the 2D image plane. Furthermore, the corresponding PV map elements are cropped and filtered according to the depths and range of images.

C. Additional Experiments

C.1. More Ablation Studies

In this section, we demonstrate more ablations on the selection of hyper-parameters. All experiments are conducted with the utilization of mimic query distillation.

Ablations on the Uncertainty Loss Weight. In Table A3, we conduct ablations on the selection of λ_{nll} for uncertainty head loss. It can be observed that the best performance is achieved when the hyperparameter λ_{nll} is set to 0.05. Excessive or insufficient weights may cause imbalances in learning, thereby affecting the model's performance.

λ_{nll}	AP_{ped}	AP_{div}	AP_{bou}	mAP
0.00	39.2	30.6	44.4	38.1
0.02	40.7	30.7	45.0	38.8
0.05	40.3	30.8	46.8	**39.3**
0.07	40.1	29.8	45.1	38.9
0.10	39.5	30.1	45.3	38.3

Table A3. Ablations on the λ_{nll} of L_{nll} for UA-Head output.

Ablations on Threshold Selection. To examine the selection scheme for PV instances, we conducted ablations on the various settings of c_{thr} in our UI2DPrompt design. As an uncertainty output, a larger c_{thr} indicates higher selection standards. Excessive or insufficient threshold selection may lead to the loss of critical features or the introduction of redundant information. As shown in Table A4, the best performance of 39.3 mAP is obtained with $c_{\text{thr}} = 0.4$.

c_{thr}	0.2	0.4	0.5	0.6
mAP	39.0	**39.3**	38.8	38.6

Table A4. Ablations on threshold selection.

D. Visualization

D.1. Visual Comparisons

On region-based splits, Figure A1 and Figure A2 present additional visual comparisons on nuScenes validation set. In Figure A1, uncertainties are presented as circles of different sizes. The larger circle represents a lower confidence in its prediction. As shown in the fifth sample, the larger uncertainty is observed in the `FRONT-RIGHT` view of the image, mainly due to occlusions caused by the car. Compared to the previous method, our UIGenMap performs better with lower uncertainties, particularly for road boundaries and pedestrian crossings. Figure A3 and Figure A4 present more qualitative visual comparisons on the region-based Argoverse2 dataset.

D.2. Scene-level Video Demo

We further provide a **video demo** named "`demo.mp4`" to demonstrate the performance of our model at the scene level. The video contains several typical driving scenes of nuScenes and Argoverse datasets.

D.3. Failure Case Analysis

Despite having greatly improved the quality of generalizable HD map construction, both the visual and numerical results show that there is still a large gap in the requirement for real-world deployment. In Figure A5, we provide some visual examples of failure cases.

Occlusion. As shown in Figure A5 (a), static map elements can be repeatedly occluded by dynamic objects on the road, which may cause a limited field of view and inadequate detection of key elements of the road.

Ambiguous Annotations. Figure A5 (b) presents an example of annotation errors. A left-turn junction can be seen on the front- and front-left-view of the PV images, which is unlabeled in the ground truth.

Low-light Conditions. For night driving and other low-light conditions, PV images cannot provide enough semantic and structural information. As shown in Figure A5 (c), it is hard to capture detailed road structures, so there is potential for further enhancement.

E. Limitations and Future Work

Limitations. Considering the generalization capability, the performance of learned models is heavily affected by the scale and diversity of the training data. In this paper, experiments are conducted within the same dataset. So, there is a lack of extension to generalization studies across different datasets. More strategies like modality fusion, data augmentation, and different modeling strategies for map element representation can be utilized for stronger generalization capability. Within the limited training data, performance on unseen driving scenarios remains constrained, which poses significant challenges for deployment in real-world applications.

Future work. In the future, further explorations are required to improve the generalization of map construction, particularly under adverse conditions such as rain, clouds,

and fog. In addition, the impact of different sensor models and placements on generalization must be addressed. For practical industrial applications, it is crucial to develop larger datasets and benchmarks that include a wide range of locations and scenarios. These resources would enable models to better manage the variability and complexity of real-world driving conditions.

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*, pages 11621–11631, 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] Adam Lilja, Junsheng Fu, Erik Stenborg, and Lars Hammarstrand. Localization is all you evaluate: Data leakage in online mapping datasets and how to fix it. In *CVPR*, pages 22150–22159, 2024. 1

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1

[5] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2021. 1

[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

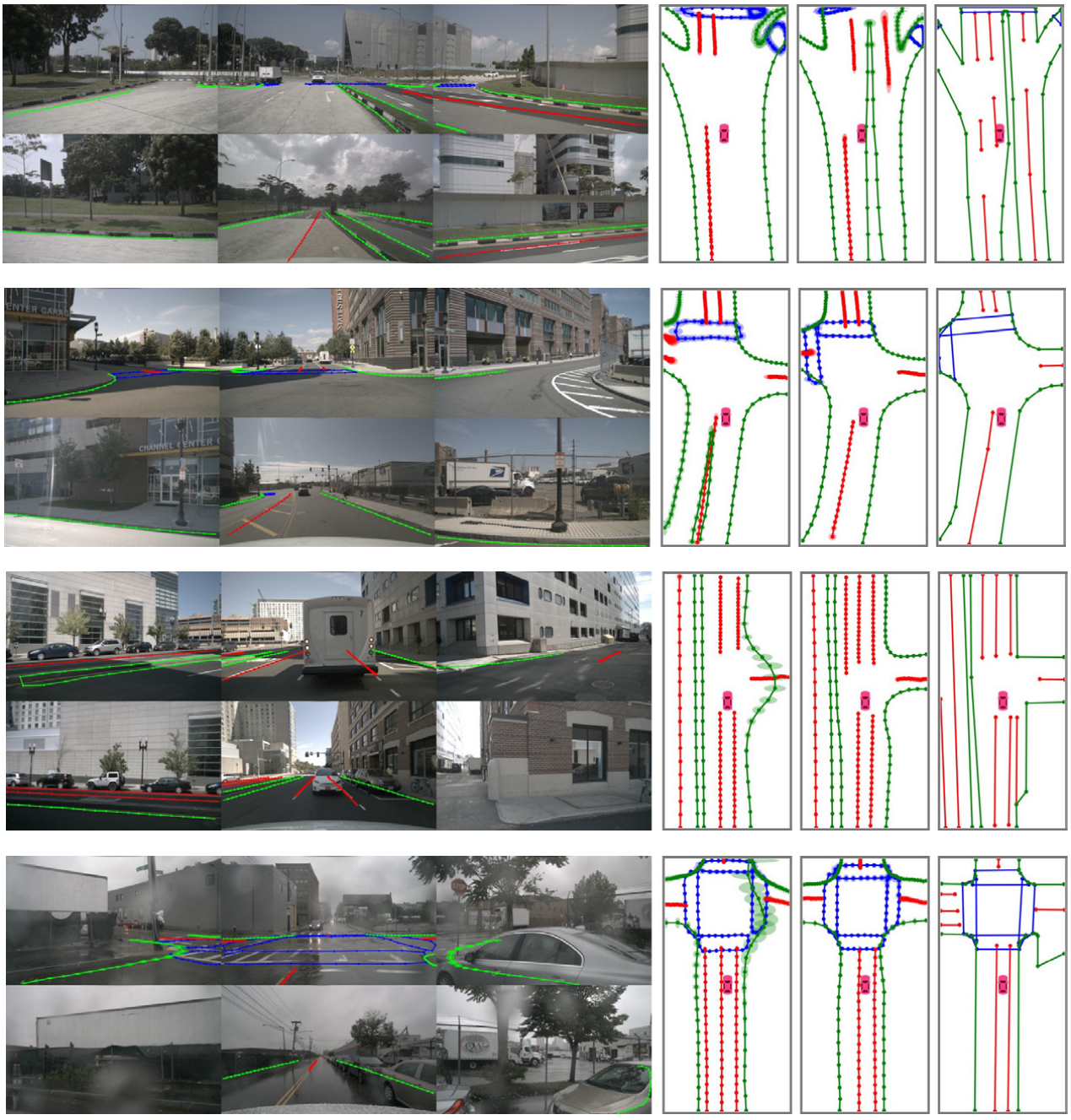


Figure A1. Visual Comparisons on region-based nuScenes validation set with PV detection result. Circles represents the point-based uncertainty, larger circle means less confidence for model's prediction. Our approach achieves better performance with less uncertainty.

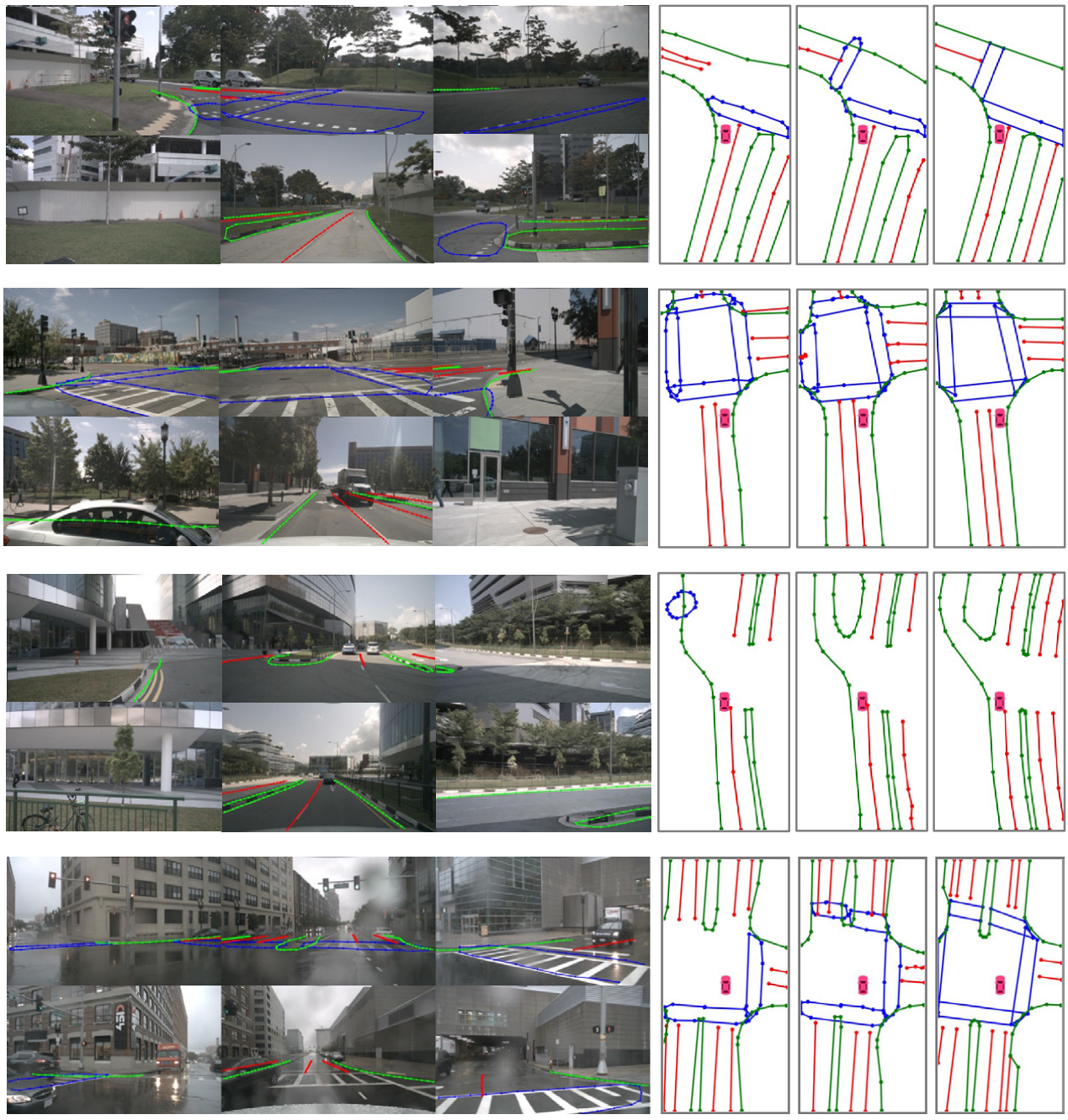


Figure A2. Visual Comparisons on region-based nuScenes validation set with PV detection result, in which our approach achieves better performance.

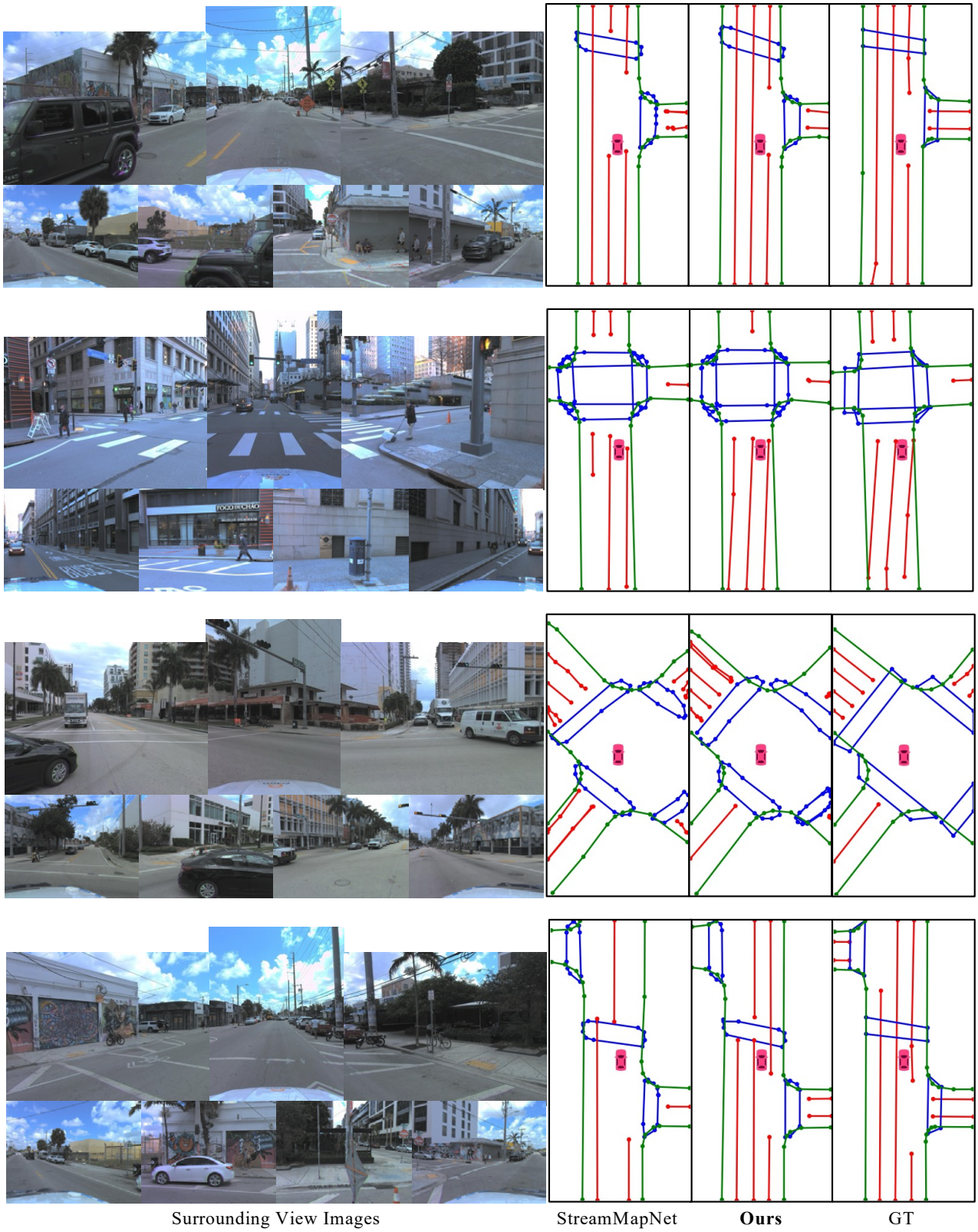


Figure A3. Visual comparisons on region-based Argoverse2 validation dataset. Our UIGenMap emphasizes stronger generalized ability.

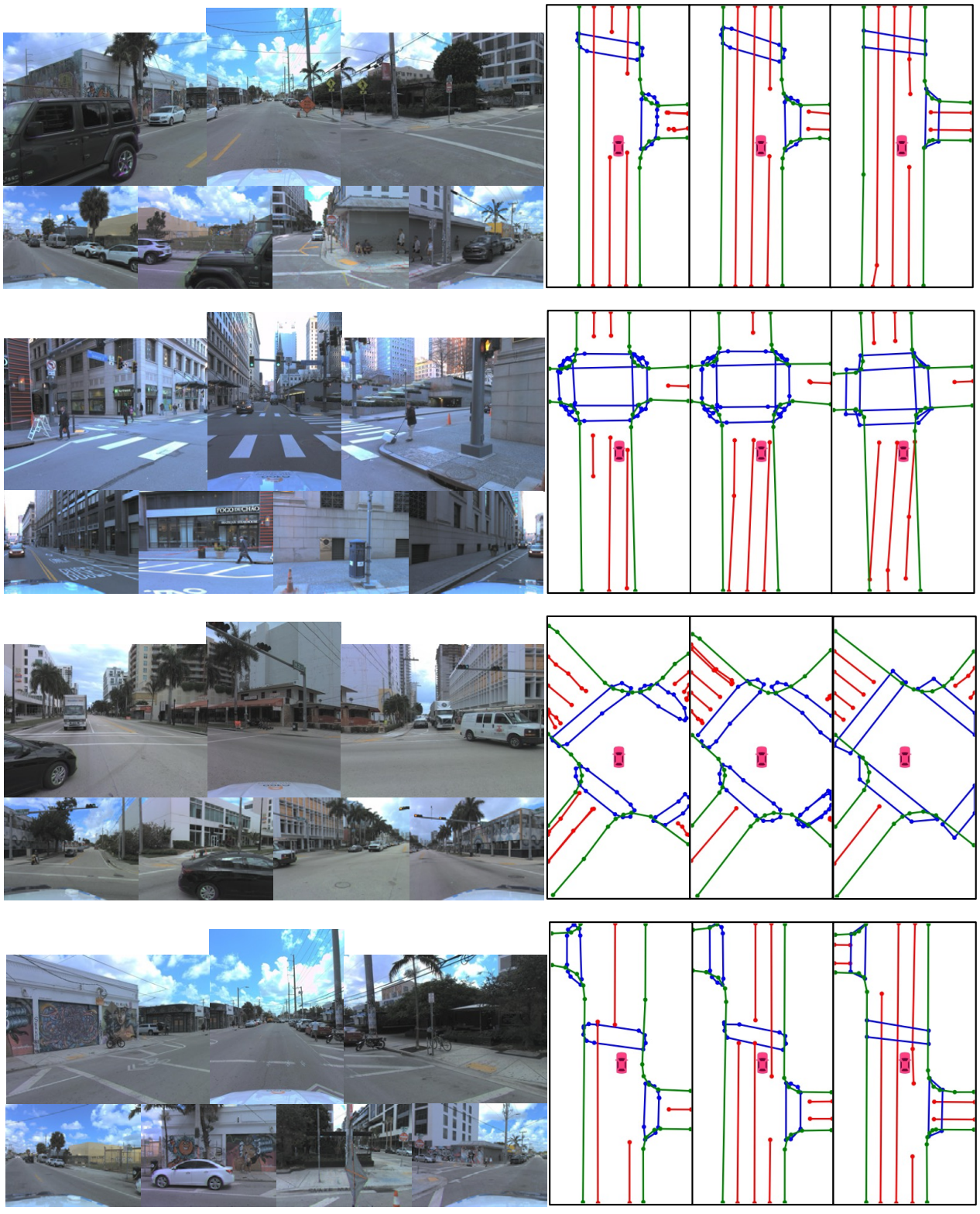


Figure A4. More visual comparisons on region-based Argoverse2 validation dataset.

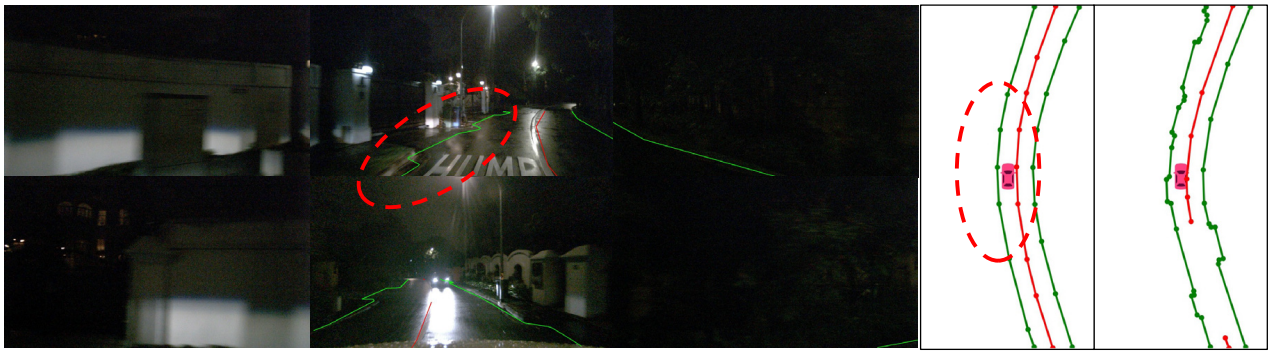
Failure Case (a): Occlusion

Surrounding Views with PV annotations | Pred | GT

Failure Case (b): Ambiguity in Annotation

Surrounding Views with PV annotations | Pred | GT

Failure Case (c): Low-light Condition



Surrounding Views with PV annotations | Pred | GT

Figure A5. Visualization of Failure cases. (a): Occlusion by dynamic objects; (b): Wrong calibration and annotations in pubulically used dataset; (c): Driving at night with low-light condition.