HOLO: Homography-Guided Pose Estimator Network for Fine-Grained Visual Localization on SD Maps

Supplementary Material

In the supplementary materials, we elaborate on the following five components to further support our paper:

A Additional details about HOLO-CA.
B Ablation of loss functions.
C Results on the Argoverse Dataset.
D More detailed run time analysis.
E Qualitative visualization.

A. Additional Details about HOLO-CA

Structure illustration. As shown in Fig. 4, given a pair of semantic features $(\mathbf{\Phi}_{bev}, \mathbf{\Phi}_{map})$, the HOLO-CA module aims to estimate the relative pose through attention-based feature fusion. We first generate patch embeddings by applying a 4×4 convolution with stride 4 to downsample the input images, yielding the patch-level feature maps $\mathbf{F}_{bev}$ and $\mathbf{F}_{map}$.

To model intra- and inter-feature interactions, We adopt self-attention and cross-attention respectively. Given BEV and map features $\mathbf{F}_{bev}$ and $\mathbf{F}_{map}$, they first undergo self-attention within each modality, followed by cross-attention between the two modalities. For an input sequence X, the self-attention and cross-attention output is computed as

$$\text{selfAttn}(X) = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) V_i, \tag{10}$$

$$\text{crossAttn}(X) = \text{softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d}}\right) V_j, \tag{11}$$

where Q, K, and V denote the query, key, and value matrices, respectively. i, j indicate different modality. In this way, BEV patches query map features, and map patches query BEV features, allowing the network to learn correspondences across the two modalities effectively.

Finally, the attention-enhanced features are concatenated and passed to the subsequent pose estimation layers. The pose estimation layers act as a decoder that transforms the fused features into either corner displacements for homography estimation or a direct 3-DoF pose.

If the network regresses corner displacements, the updated four corner points parameterize a perspective transform used to warp input BEV features $\mathbf{\Phi}_{bev}$:

$$\mathbf{\Phi}_{bev}^{(t+1)} = \mathcal{H}(\mathbf{p}^{(t+1)}) \cdot \mathbf{\Phi}_{bev}^{(t)},$$

where $\mathcal{H}$ denotes a differentiable homography estimator. The warped image is re-encoded at the next iteration, enabling coarse-to-fine refinement.

Implementation Details. The input semantic features $\mathbf{\Phi}_{bev}$ and $\mathbf{\Phi}_{map}$ share the same spatial resolution as those fed into the HOLO, while the patch embeddings $\mathbf{F}_{bev}$ and $\mathbf{F}_{map}$ are of size $64 \times 64 \times 256$. During feature fusion, we compress each patch token from 256 dimensions to 96 dimensions for constructing the query, key, and value vectors. We use a three-layers attention block.

Table 7. Comparison of localization accuracy across different numbers of iterations.

Iteration	Position Recall@Xm ↑				Orientation Recall@$X°$ ↑			
	$1m$	$2m$	$5m$	$10m$	$1°$	$2°$	$5°$	$10°$
1	21.47	46.70	77.71	90.02	39.52	76.13	94.16	98.41
2	23.49	49.60	79.00	90.52	39.52	74.27	94.96	99.20

Table 8. Comparison of the model performance across different numbers of iterations.

Iteration	GFLOPs ↓	FPS ↑
1	109.10	23.47
2	118.93	16.41

Additional Experiments. We evaluate the localization accuracy and model performance of HOLO-CA with homography head under different iteration numbers. Increasing the number of iterations from 1 to 2 leads to a moderate improvement in localization accuracy. However, unlike HOLO, HOLO-CA must recompute the feature fusion module at every iteration. As shown in Tab. 8, this iterative recomputation leads to rapidly escalating computational cost, which in turn causes a pronounced degradation in inference speed.

B. Ablation of Loss Functions

To construct feature pairs with homography, we introduce BEV and map semantic losses to encourage semantic alignment between the two modalities. The detailed expression of semantic loss $\mathcal{L}_{\text{sem}}$ is

$$\mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{BEV}}^{\text{sem}} + \mathcal{L}_{Map}^{\text{sem}}. \tag{12}$$

As shown in Tab. 9, adding the BEV semantic loss yields a substantial improvement in localization accuracy, indicating that this loss helps bring BEV features closer to map features in both geometry and semantics. With the additional map semantic loss, the accuracy is further improved,

Table 9. Localization results of various combination of loss functions on nuScenes dataset.

Pose Loss	BEV Loss	Map Loss	Recall@$Xm\uparrow$				Orientation Recall@$X°\uparrow$				APE$(m)\downarrow$	AOE$(°)\downarrow$
			$1m$	$2m$	$5m$	$10m$	$1°$	$2°$	$5°$	$10°$		
✓	✓	✓	**36.41**	**61.21**	**84.10**	**93.09**	**73.74**	**91.51**	97.35	99.20	**3.37**	**1.02**
✓	✓		30.49	58.70	83.01	92.77	71.88	91.51	**98.14**	99.20	3.63	1.04
✓			17.01	40.91	77.82	90.72	56.50	85.68	97.61	**99.47**	4.58	1.31

suggesting that it effectively narrows the modality gap and provides higher-quality paired features.

C. Results on the Argoverse Dataset.

We further conduct experiments on the Argoverse dataset. Argoverse is also an autonomous driving dataset collected in Miami and Pittsburgh. Following the official split, we use 13,122 samples for training and 5,015 samples for validation. Similar to nuScenes, Argoverse does not provide SD map data; therefore, we obtain the corresponding SD maps from OpenStreetMap (OSM). In contrast to nuScenes, this dataset does not require any alignment.

Table 10. Localization results on argoverse dataset. * denotes results directly reported in the original paper, where the model is first trained on nuScenes and then fine-tuned on Argoverse.

Method	Position Recall@$Xm\uparrow$			Orientation Recall@$X°\uparrow$		
	$1m$	$2m$	$5m$	$1°$	$2°$	$5°$
MapLocNet*	23.26	47.24	79.13	62.35	86.28	96.24
HOLO	29.38	51.89	77.26	65.61	86.33	95.86

Tab. 10 reports the performance of our method on the Argoverse dataset. Since the results of MapLocNet are obtained by first training on nuScenes and then fine-tuning on Argoverse, whereas our model is trained from scratch directly on Argoverse, with significantly fewer training samples, our Recall@5m/5° is slightly lower. However, our method substantially outperforms MapLocNet on Recall@1m/2m, demonstrating the superior localization accuracy of our approach.

D. More Detailed Run Time Analysis

Fig. 6 presents the detailed inference-time breakdown of HOLO under the 6-iteration setting. Overall, the frontend feature perception stage and the backend pose estimation stage each account for roughly half of the total runtime, indicating a balanced computational load between the two. Within the feature perception stage, the image backbone and the BEV view transformation dominate the computation cost, representing the primary bottlenecks of the pipeline. These observations suggest that further optimization of image encoding and BEV generation will be crucial for improving the overall efficiency of the system in real-world deployment.

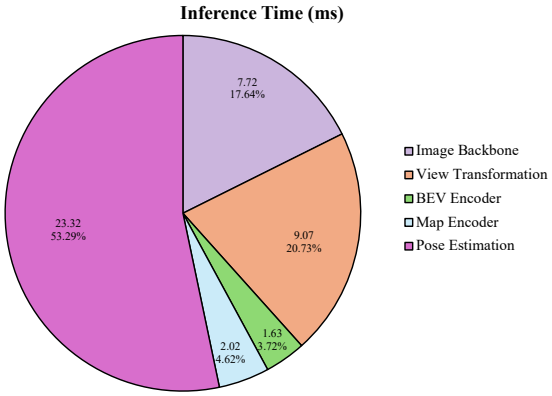

Figure 6. Detailed run time. We conducted an inference time analysis of each component of HOLO on an NVIDIA RTX A6000 GPU.

E. Qualitative Visualization

Segmentation Visualization. Fig. 7 shows the semantic segmentation results of the BEV perception module when supervised by the drivable areas provided by nuScenes and the road elements extracted from OSM, respectively. Each row in the figure corresponds to segmentation results at the same physical location. As illustrated by the second row, the road annotations in the SD map deviate from the actual road structure observed by the camera. The drivable-area ground truth indicates the presence of a fork at that location, whereas the SD map labels fail to provide this information. Consequently, the BEV perception module's predictions become misaligned with the SD map, degrading localization accuracy. In contrast, supervision from drivable areas offers annotations that closely match real-world observations and contain richer structural details, enabling more accurate localization.

Map Alignment Visualization. Based on the procedure for converting WGS84 coordinates into the ENU coordinate system, the alignment results between our SD map and the nuScenes HD map are shown in Fig. 8. The road geometry from the SD map (red curves) aligns well with the drivable-

area layer of the HD map (light blue regions). However, due to the limited positional accuracy of the SD map, certain misalignments still occur. In addition, because the SD map is not updated in a timely manner, some road segments that appear in the HD map are missing from the SD map. Fig. 9 presents the alignment results between the Argoverse HD map and our collected SD map. Since the original map and the SD map share the same coordinate system, no additional alignment procedure is required.

Qualitative Results. Fig. 10 further presents additional qualitative results of HOLO on the nuScenes dataset. Beyond typical daytime scenes, the fourth to sixth rows illustrate localization performance under more challenging conditions such as rainy weather and nighttime driving. These results demonstrate that our method maintains stable localization accuracy despite variations in illumination, adverse weather, and other forms of visual degradation, highlighting its strong robustness across diverse environments.

Failure Case Study. As shown in Fig. 11, we also examine several representative failure cases of the model's localization results, which may offer insights for improving future approaches and inspire further research in this direction.

Case 1. The first category of large localization errors primarily arises from the sparsity of discriminative features along the road direction. As shown in the visualization, both the BEV features and the map features exhibit very limited variation along the longitudinal direction of the road. This lack of distinctive cues provides insufficient geometric constraints for the model, leading to significant localization drift along the road axis. In contrast, the features vary more prominently in the direction perpendicular to the road, allowing the model to establish stronger constraints and achieve higher localization accuracy in that dimension.

Case 2. The second source of localization error is the presence of dense traffic. Since our homography estimation relies primarily on static landmarks such as road boundaries and building contours, heavy traffic can significantly degrade the quality of BEV perception. As illustrated in the figure, densely packed vehicles obscure large portions of the roadway, resulting in blurred or incomplete road structures with weak geometric cues. This degradation makes it more difficult to estimate a reliable homography, ultimately leading to increased localization errors.

Case 3. The third type of large localization error arises from the presence of many structurally similar regions in the SD map. Because the BEV perception has a limited field of view while the SD map typically covers a much larger area, the model may encounter multiple regions in the map that share highly similar local patterns with the observed BEV features. This often leads to ambiguous associations and incorrect matches. In the illustrated example, the local geometric structure around the ground-truth position closely resembles that of the model's estimated position, making it difficult for the network to disambiguate between the two and ultimately resulting in a significant localization offset.

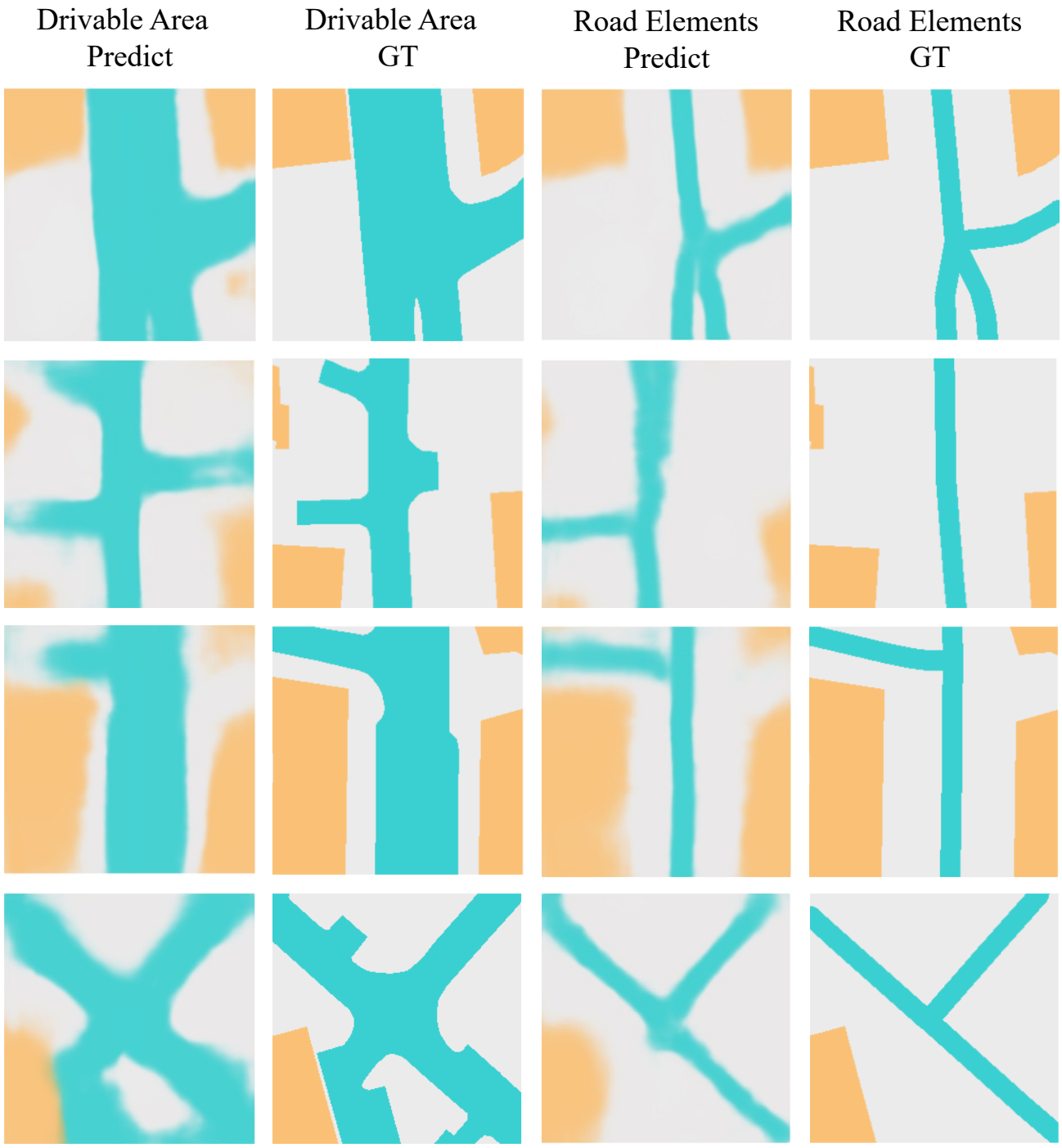

Figure 7. The segmentation results of the BEV perception module under the supervision of drivable area and road elements, respectively.

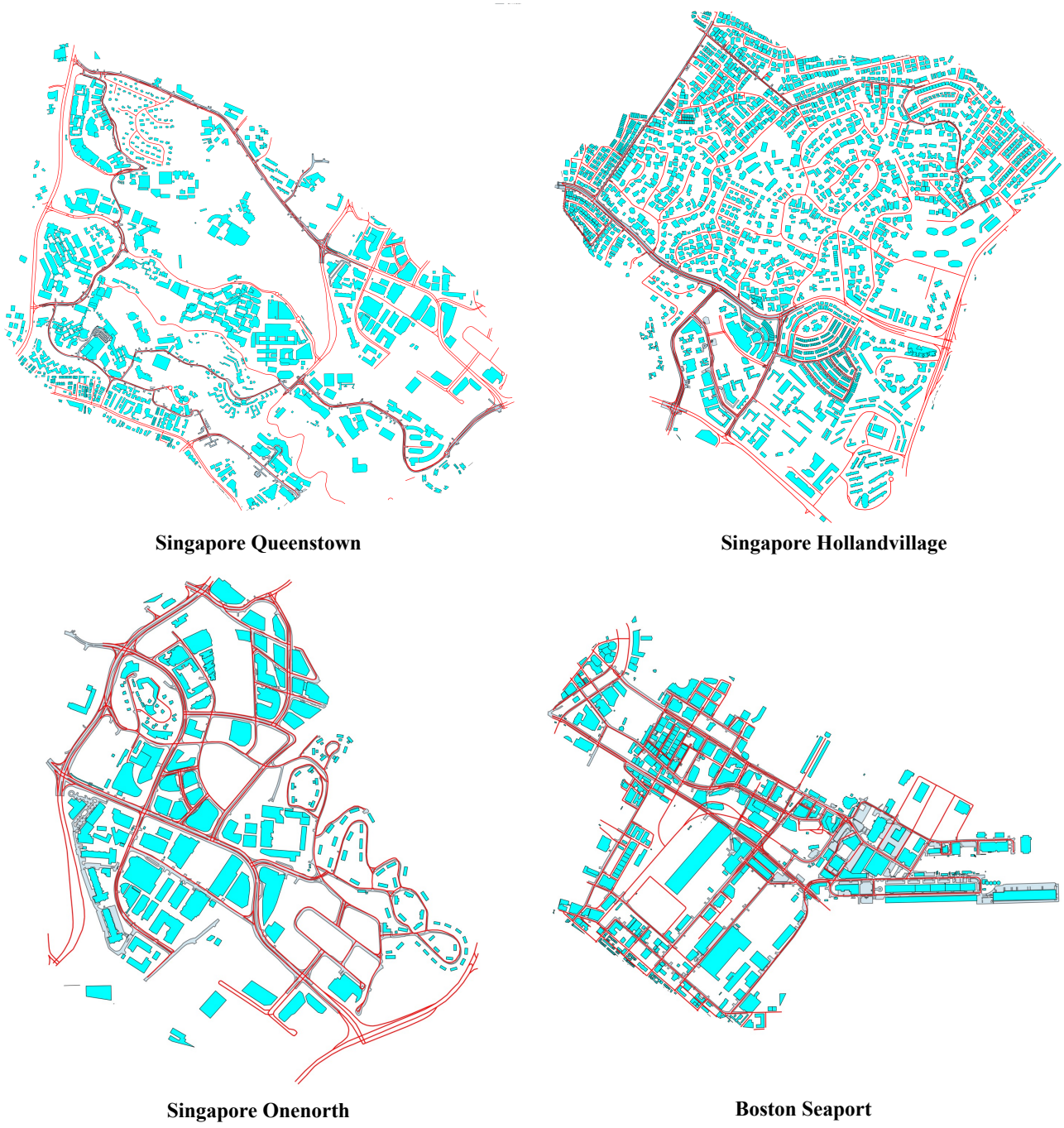

Figure 8. The alignment results between the nuScenes HD map and the OSM map collected in our study. The red curves represent the roads extracted from the SD map, while the light blue regions correspond to the drivable-area layer in the HD map. The cyan regions indicate the building areas in the SD map.

Figure 9. The alignment results between the Argoverse HD map and the OSM map collected in our study. The red curves represent the roads extracted from the SD map, while the light blue regions correspond to the drivable-area layer in the HD map. The cyan regions indicate the building areas in the SD map.

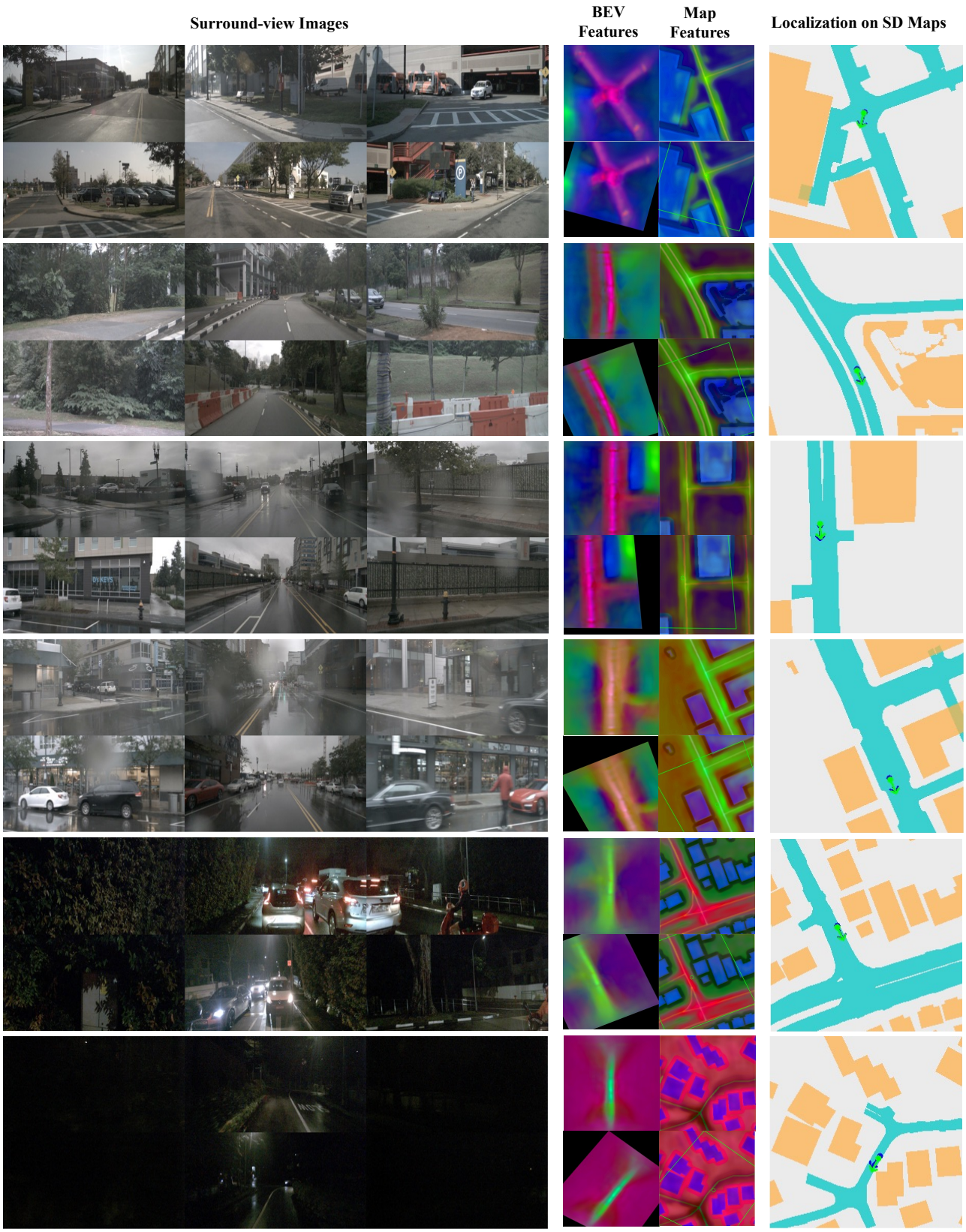

Figure 10. Additional qualitative results for visual localization on nuScene dataset.

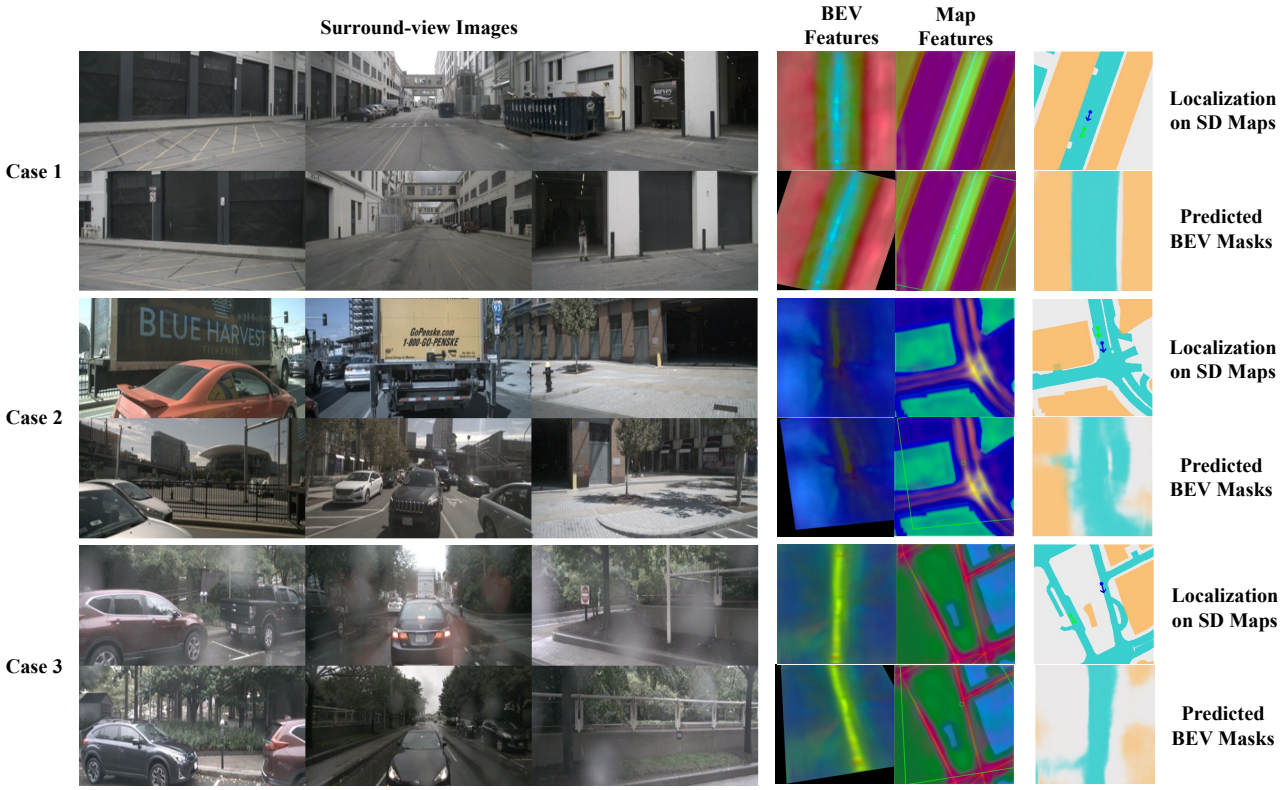

Figure 11. Visualization of three typical failure cases in our localization framework.