U-ViLAR: Uncertainty-Aware Visual Localization for Autonomous Driving via Differentiable Association and Registration

1. Overview

In addition to the main paper, we provide supplementary experimental results, detailed insights into our real-world road test scenario dataset, comprehensive qualitative analyses, and an expanded summary to enhance understanding.

This is a brief overview:

- section 2: Comprehensive Experimental Results
- section 3: SRoad Dataset
- section 4: Visualization

2. Comprehensive Experimental Results

In the following sections, we will delve into the details not thoroughly covered in the main text.

2.1. HD Map.

The high-definition map (HD Map) in the nuScenes dataset typically includes the following elements and precision information:

- **Road Boundaries**: Define the edges of the road, including shoulders and medians.
- **Lane Dividers**: Markings used to distinguish different lanes, such as dashed or solid lines.
- **Lane Centerlines**: Central reference lines of lanes, used for navigation and localization.
- **Crosswalks**: Markings for pedestrian crossing areas.
- **Traffic Signs**: Including traffic lights, stop signs, etc.
- **Road Markings**: Ground markings such as arrows and text.

However, when actually inputting into the network, we only utilize information related to the road structure.

2.2. Navigation Map.

OpenStreetMap (OSM) is a collaborative project to create a free and editable world map. It was founded by Steve Coast in 2004 as a response to the lack of freely available geographic data from national mapping agencies like the Ordnance Survey in the UK. OSM allows volunteers to contribute data through surveys, GPS traces, aerial imagery, and other freely licensed sources. The project is maintained by the OpenStreetMap Foundation and is licensed under the Open Database License, making it widely used for electronic maps, navigation, humanitarian aid, and data visualization.

OSM's data structure consists of three primary elements:

- **Nodes**: Points defined by latitude and longitude, representing specific locations such as landmarks or buildings.
- **Ways**: Lines or areas formed by connecting nodes, used to represent roads, rivers, or boundaries. Ways can be open polylines, closed polylines, or areas.
- **Relations**: Groups of elements (nodes, ways, or other relations) that describe relationships, such as routes, administrative boundaries, or restrictions.

Additionally, **Tags** (key-value pairs) are used to describe the attributes of these elements. For example, `highway=residential` defines a residential road, while `maxspeed:winter=⋆` specifies winter speed limits. In the actual input to the network, similar to Orienternet, the OSM data is rendered at a resolution of 0.5 meters per pixel, providing detailed vector maps with 48 types of semantic information, including road networks and building footprints.

2.3. Implementation Details.

The localization framework processes $[224 \times 400]$ input images through ResNet-18 to generate 32-channel BEV (Bird's Eye View) features, with task-specific configurations designed to address different localization challenges. For fine-grained localization, a high-resolution perception range of $[-60\text{m}, 60\text{m}]$ longitudinally and $[-15\text{m}, 15\text{m}]$ laterally is established, where the resolution is set to 5 me-

Methods	Inputs	Lateral Error ↓		Longitudinal Error ↓		Orientation Error ↓	
		MAE(m)	RMSE(m)	MAE(m)	RMSE(m)	MAE(°)	RMSE(°)
Ours-M	SRoad + HD map	0.110	0.136	0.284	0.322	0.090	0.124
+ sequence	SRoad + HD map	0.086	0.102	0.205	0.245	0.065	0.089
+ uncertainty-aware sequence	SRoad + HD map	**0.076**	**0.090**	**0.182**	**0.209**	**0.059**	**0.074**

Table 1. **Our ablation experiments on temporal localization using HD maps on the SRoad dataset with our single-frame model.**

Methods	Inputs	Lateral Error ↓		Longitudinal Error ↓		Orientation Error ↓	
		MAE(m)	RMSE(m)	MAE(m)	RMSE(m)	MAE(°)	RMSE(°)
Ours-M	nuScenes + HD map	0.040	0.049	0.140	0.158	0.075	0.089
w/o PDP.	nuScenes + HD map	0.042	0.054	0.144	0.184	0.079	0.099
Ours-M	SRoad + HD map	0.110	0.136	0.284	0.322	0.090	0.124
w/o PDP.	SRoad + HD map	0.125	0.169	0.305	0.398	0.105	0.148

Table 2. **Ablation experiments of Pose Distribution Prior of LU-Guided Registration on the nuScene and SRoad.**

ters per pixel (5m/px). To simulate GPS noise, HD (High-Definition) maps are perturbed with small random transformations, including rotational perturbations within the range of $\theta \in [-2^\circ, 2^\circ]$ and translational perturbations within the range of $t \in [-2\text{m}, 2\text{m}]$. Following this, a 120m×120m search area with a resolution of 5m/px is extracted, centered on the ego vehicle, to facilitate precise localization.

Conversely, for the task of relocalization, a coarser resolution of 8 meters per pixel (8m/px) is employed, covering a larger area of $[-64\text{m}, 64\text{m}]$ longitudinally and $[-32\text{m}, 32\text{m}]$ laterally. To address significant deviations that may occur in relocalization scenarios, larger perturbations are introduced, including rotational perturbations within the range of $\theta \in [-30^\circ, 30^\circ]$ and translational perturbations within the range of $t \in [-30\text{m}, 30\text{m}]$. The framework then processes a 128m×128m search region with a resolution of 8m/px on the navigation map to ensure robust relocalization.

In the local association constraint, anchor points are established at intervals of 5 meters laterally and 10 meters longitudinally to provide a structured reference for localization. Additionally, associated point pairs are constructed within a window of [6m, 12m], ensuring that the localization framework can effectively associate and match points within this predefined spatial range. This structured approach enhances the accuracy and reliability of the localization process across both fine-grained and relocalization tasks.

We train the model using 8 NVIDIA V100 GPUs for 160 epochs, which takes approximately 36 hours to converge. The model is optimized using an AdamW optimizer with a weight decay of 1e-4, a batch size of 8, and an initial learning rate of 1e-4. During training, we employ a cosine annealing scheduler to adjust the learning rate dynamically.

2.4. ICP-based Method.

We have chosen a non-learning-based Iterative Closest Point (ICP) method as the benchmark for rule-based approaches. ICP is a classic algorithm widely used for aligning geometric or semantic features, whose core idea is to iteratively optimize the correspondence between observed visual elements (e.g., lane markings, stop lines, and road boundaries) and their semantic counterparts in high-definition maps, thereby estimating the relative pose (position and orientation) between them. Specifically, the ICP method begins by sampling key visual features and their semantic equivalents in the map and establishing correspondences through nearest feature association. It then optimizes the pose transformation by minimizing the distance between matched features until convergence. In certain cases, the ICP method can be combined with the Umeyama algorithm, leveraging its closed-form solution to rapidly estimate rotation and translation transformations, further enhancing computational efficiency.

The reason for selecting ICP as the benchmark method lies in its status as a classic algorithm in feature alignment, enjoying widespread industry recognition and mature application foundations. As a fundamental algorithm in the field, ICP demonstrates strong robustness and versatility across various scenarios, particularly in providing stable pose estimation results without relying on data-driven models. By comparing with the ICP method, we can effectively evaluate the performance, accuracy, and efficiency of other approaches (e.g., learning-based methods), while providing a reliable reference standard for subsequent research. Additionally, the transparency and interpretability of the ICP method make it an ideal choice for validating the effectiveness of new algorithms, especially in scenarios where visual and semantic elements are matched without the use of point clouds.

PU-Guided Association	LU-Guided Registration	Lateral Recall@Xm ↑			Longitudinal Recall@Xm ↑			Orientation Recall@X° ↑		
		1m	3m	5m	1m	3m	5m	1°	3°	5°
✓	✓	69.12	91.25	93.68	32.04	63.00	70.20	64.92	94.84	97.44
	✓	52.32	85.23	92.01	26.32	54.54	67.00	42.54	76.92	86.95
✓		55.95	87.53	93.92	29.44	58.88	69.23	53.33	85.23	93.48
		50.23	83.95	91.90	24.96	53.48	66.01	30.23	70.53	82.54

Table 3. **Ablation of PU-Guided Association and LU-Guided Registration using KITTI and OSM.**

2.5. Enhancing Temporal Filtering with Localization Uncertainty

To demonstrate that the 3DoF visual localization uncertainty output can significantly improve the accuracy and robustness of temporal filtering results, we designed the following experimental framework. Specifically, we incorporate localization uncertainty as the variance of observation noise into the filtering model during the filtering process. When using a Kalman filter for temporal filtering, the uncertainty dynamically adjusts the observation noise covariance matrix R_k in the update step. The implementation details are as follows:

Uncertainty-Aware Filtering Mechanism First, we define the observation noise covariance matrix R_k to be proportional to the localization uncertainty:

$$R_k = \alpha U_k. \tag{1}$$

where U_k represents the current frame's localization uncertainty, and α is a scaling factor that modulates the uncertainty's impact on filtering.

The prediction step utilizes the previous state estimate and state transition model:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + Bu_k \tag{2}$$

$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q. \tag{3}$$

In the update step, the Kalman gain adapts to the uncertainty-adjusted R_k:

$$K_k = P_{k|k-1}H^T(HP_{k|k-1}H^T + R_k)^{-1}. \tag{4}$$

Subsequent state correction becomes:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H\hat{x}_{k|k-1}) \tag{5}$$

$$P_{k|k} = (I - K_kH)P_{k|k-1}. \tag{6}$$

This adaptive mechanism ensures that higher uncertainty values (U_k) induce larger observation noise covariance (R_k), thereby reducing filter confidence in current measurements and increasing reliance on prior predictions. The dynamic adjustment capability significantly enhances filter robustness during periods of elevated localization uncertainty, particularly in challenging scenarios with reduced longitudinal constraints. Quantitative validation of this improvement is presented below.

Results. As shown in Table 1, the baseline method Ours-M, using SRoad data and HD maps, significantly reduces all error metrics when introducing sequential filtering. Furthermore, the sequential filtering enhanced with localization uncertainty demonstrates even lower errors.

2.6. Ablation of Pose Uncertainty Prior on SRoad.

In Section 4.3 of the main text, we concluded that omitting the Pose Uncertainty Prior as input information for registration leads to a certain degree of deterioration in our primary metric, MAE, reflecting a decline in localization performance. Notably, RMSE experiences an even more pronounced degradation. This result aligns with the fundamental design purpose of the Pose Uncertainty Prior, which is specifically intended to be effective in the less common non-unimodal scenarios within the dataset. We conducted ablation experiments on the more challenging SRoad dataset to strengthen this conclusion. As shown in Table 2, it is evident that on SRoad, compared to nuScenes (which contains more straightforward scenarios), the omission of the Pose Uncertainty Prior results in a more significant increase in MAE error.

Ablation of PU-Guided Association and LU-Guided Registration on KITTI. In Section 4.3 of the main text, we demonstrated the importance of the relevant modules using nuScenes and HD maps in the Fine-grained Localization task. Additionally, we conducted ablation experiments on the Large-Scale Relocalization task using KITTI and OSM. As shown in Table 3, the experimental results exhibit consistent conclusions.

3. SRoad Dataset

The SRoad dataset is a comprehensive and diverse dataset specifically designed to address the challenges of modern urban driving scenarios. It spans over 30 cities and captures various road structures, traffic conditions, and environmental factors. With more than 500,000 frames, including a dedicated test set of 100,000 frames, the SRoad dataset provides

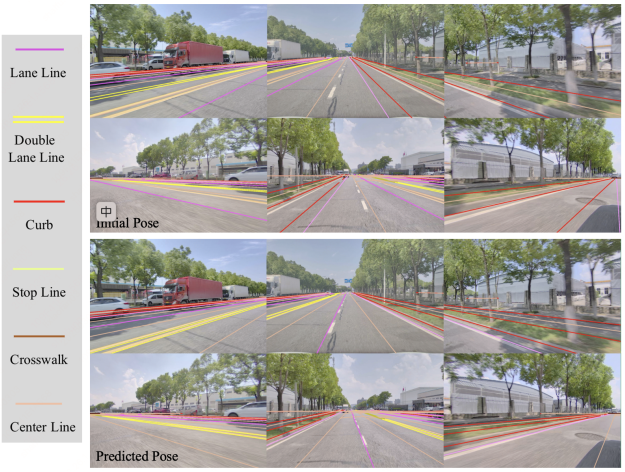

Figure 1. The effect of projecting a map onto a camera image based on the initial pose(upper) and the final results predicted by U-ViLAR(lower).

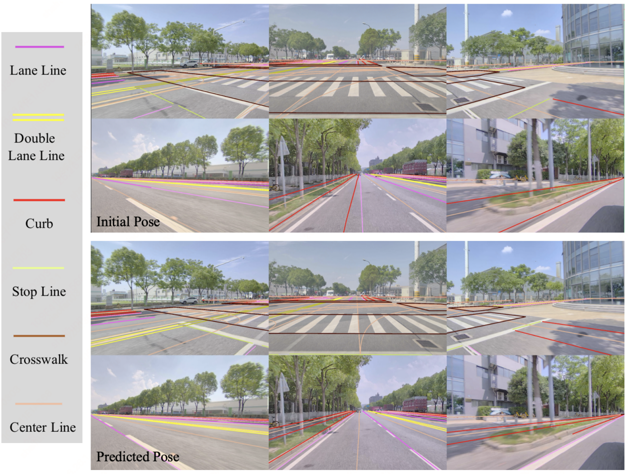

Figure 2. The effect of projecting a map onto a camera image based on the initial pose(upper) and the final results predicted by U-ViLAR(lower).

a robust foundation for developing and evaluating reliable localization and mapping algorithms. Notably, over 60% of the dataset scenarios present specific challenges critical for testing the robustness and generalization capabilities of autonomous driving systems. These challenges include complex intersections, merging and diverging zones, congested areas, and areas under viaducts—all common in real-world urban environments but often underrepresented in existing datasets. By incorporating such diverse and challenging scenarios, the SRoad dataset enables us to push the boundaries of current technologies, ensuring that our solutions are not only accurate but also adaptable to the complexities of real-world driving conditions. We discuss the selection of sensors for the SRoad dataset collection vehicle in Section 3.1, and the methods for acquiring ground truth localization are described in Section 3.2.

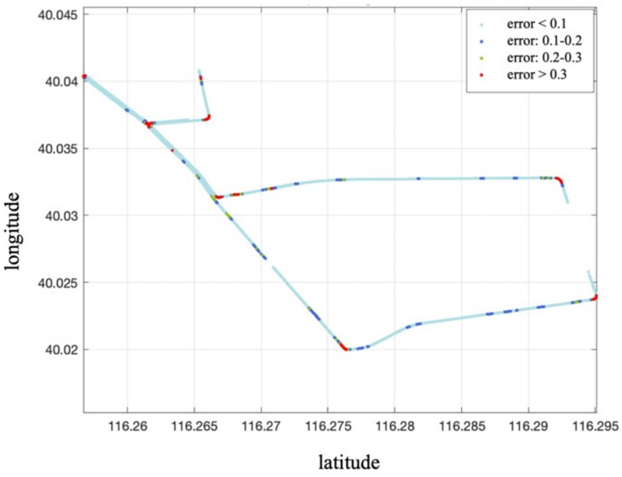

Figure 3. We plot the 3-DOF maximum absolute error of U-VilLAR's predicted pose in the test set(The lateral and longitudinal disturbance values were set as 2m, and the heading Angle disturbance values were set as 2°). The horizontal and vertical axes are latitude and longitude respectively.

3.1. Sensor Suite

The SRoad data acquisition vehicle has an extensive sensor suite, designed to capture a comprehensive range of environmental data. The sensor configuration is as follows:

- **Cameras:** The vehicle includes seven cameras, covering various perspectives. The specific models are onsemi_narrow, onsemi_obstacle, spherical_backward, spherical_left_backward, spherical_left_forward, spherical_right_backward, and spherical_right_forward. These cameras are mounted around the vehicle, offering 360-degree coverage. Please refer to the sensor parameter file we provided for specific details.
- **LiDAR System:** A 1+4 LiDAR setup is employed, consisting of a top-mounted Hesai P90 and four additional rs-bpearl LiDARs positioned for blind spot reduction at the front, rear, left, and right sides of the vehicle. (Current data collection vehicles suffer from significant blind spots. We aim to achieve over 90% coverage of these existing blind areas by integrating supplementary LiDAR sensors. Figure 4 illustrates the field of view and blind spots of the Hesai P90.)

Apart from the primary visual and LiDAR sensors, the vehicle is outfitted with a range of auxiliary sensors to enhance data accuracy and environmental understanding:

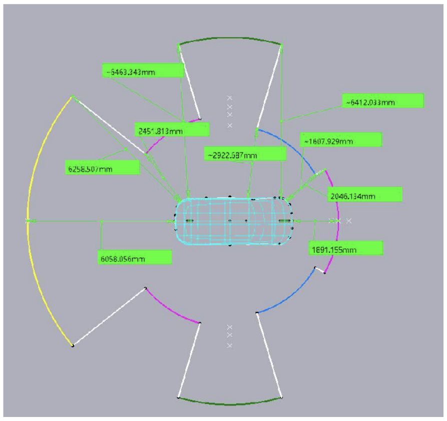

Figure 4. Field of view and blind spots of the Hesai P90 LiDAR sensor.

- **Onboard Antenna:** The GNSS-850 antenna encompasses a comprehensive range of GNSS frequencies.
- **IMU:** The ISA-100C is a near-navigation-grade IMU sensor, featuring fiber optic gyroscopes and a full-temperature-compensated Micro-Electro-Mechanical Systems (MEMS) accelerometer.
- **GNSS Receiver:** The PwrPak7D by Novatel is a dual-antenna GNSS receiver, supporting most GNSS system frequencies.
- **4G Network Connection:** The MD-649 module ensures consistent connectivity.
- **Novatel Equipment:** The PP7 device is a critical component for data acquisition.

3.2. Acquisition of Ground Truth

The PP7 device outputs GNSS positions and differential heading data, complemented by low-precision IMU measurements. The ISA-100C provides high-precision IMU measurements. These data, integrated with raw satellite observations recorded by the PP7, are processed through the IEOUT software to yield high-accuracy positioning ground truth.

4. Visualization

We provide video and image visualizations that more effectively demonstrate our method.

4.1. Video Result

Performance Video on SRoad Dataset We provide two videos in mp4 format, displaying the global localization performance under different quality initial values, as well as a comparison with traditional ICP-based methods: `Demo_U-ViLAR.mp4`

High-Precision Ground Truth Demonstration of SRoad Dataset We also provided three videos in mp4 format, including the original information collected by the collection vehicle, the calculated localization ground truth, and the reconstruction results of obstacles/road structure, demonstrating the rigor of the SRoad dataset truth value production method.

- `SRoad_localization_trajectory.mp4`
- `SRoad_obstacle_rec.mp4`
- `SRoad_road_juction_rec.mp4`

4.2. More Results Visualization

As shown in Figures 1, we compared the effect of projecting a map onto a camera image based on the initial pose(upper) and the final results predicted by U-ViLAR(lower).

As shown in Figures 3, we plot the 3-DOF maximum absolute error of U-VilLAR's predicted pose in the test set (the lateral and longitudinal disturbance values were set as 2 meters, and the heading angle disturbance values were set as 2°). The horizontal and vertical axes represent latitude and longitude, respectively. We observe that the positions with more significant errors are mostly concentrated in areas such as intersections.

We visualized the performance in scenarios where localization stability is challenging and showcased its advantages over the traditional ICP-Based method:

- Figures 5 to 7 demonstrate that both the ICP-Based Method(indicated in red) and ours(indicated in green) errors are within normal ranges. However, longitudinal observations in this scenario caused a certain degree of degradation. As a result, the longitudinal localization still remains near the initial pose estimate of GNSS (indicated in orange). It has not been optimized to the ground truth value (indicated in white). Nevertheless, ours exhibits smaller overall errors in both longitudinal and lateral directions.
- Figures 8 to 11 illustrate that the traditional ICP-Based method, due to poor detection of some lane lines, leads to matching failures and more significant lateral errors, whereas ours exhibits smaller lateral errors
- Figures 12 to 15 show that in intersection/merging and diverging scenarios, the traditional ICP-Based method experiences increased longitudinal errors due to poor detection or incorrect association of some lane lines, whereas ours has smaller errors in both lateral and longitudinal directions.

- Figures 16 to 18 demonstrate that the matching in the traditional ICP-Based method fails to compute registration values due to lane lines being detected too short or not detected at all, while ours continues to operate normally.

References

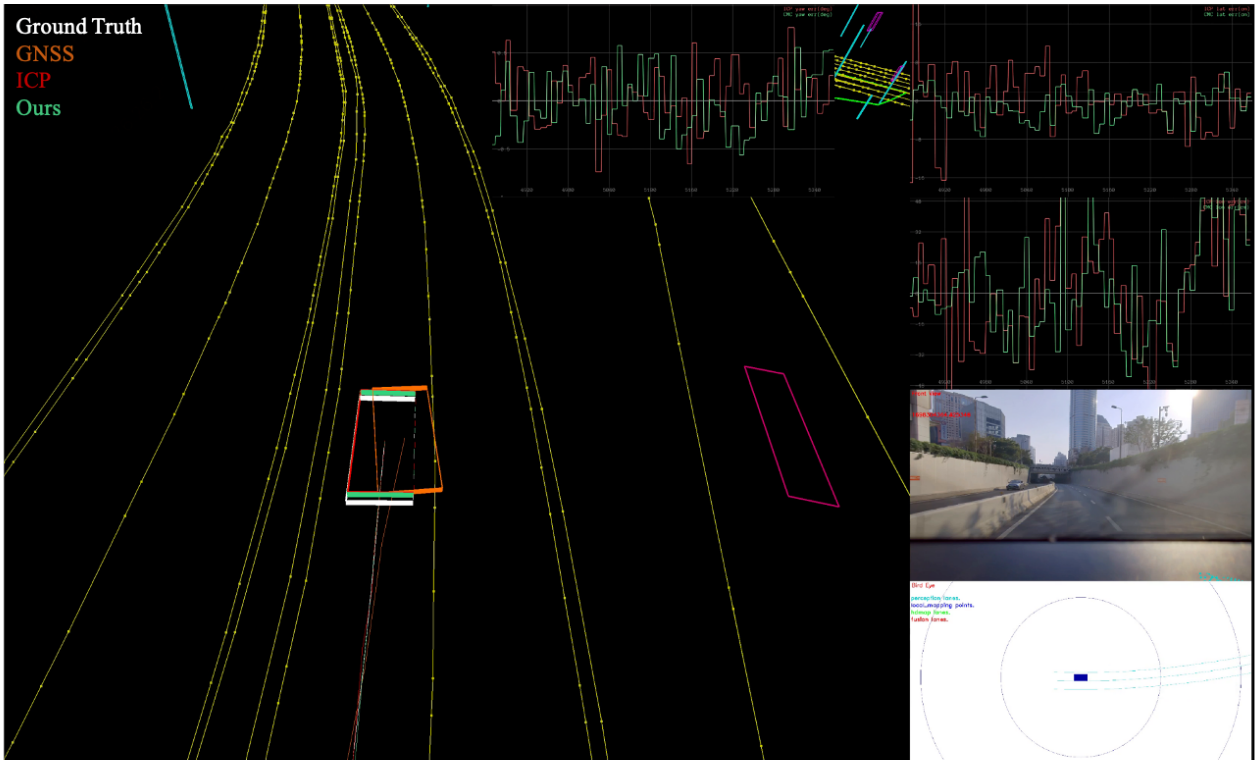

Figure 5. Visualization of Localization Performance in Scenario 1 within the SRoad Dataset.

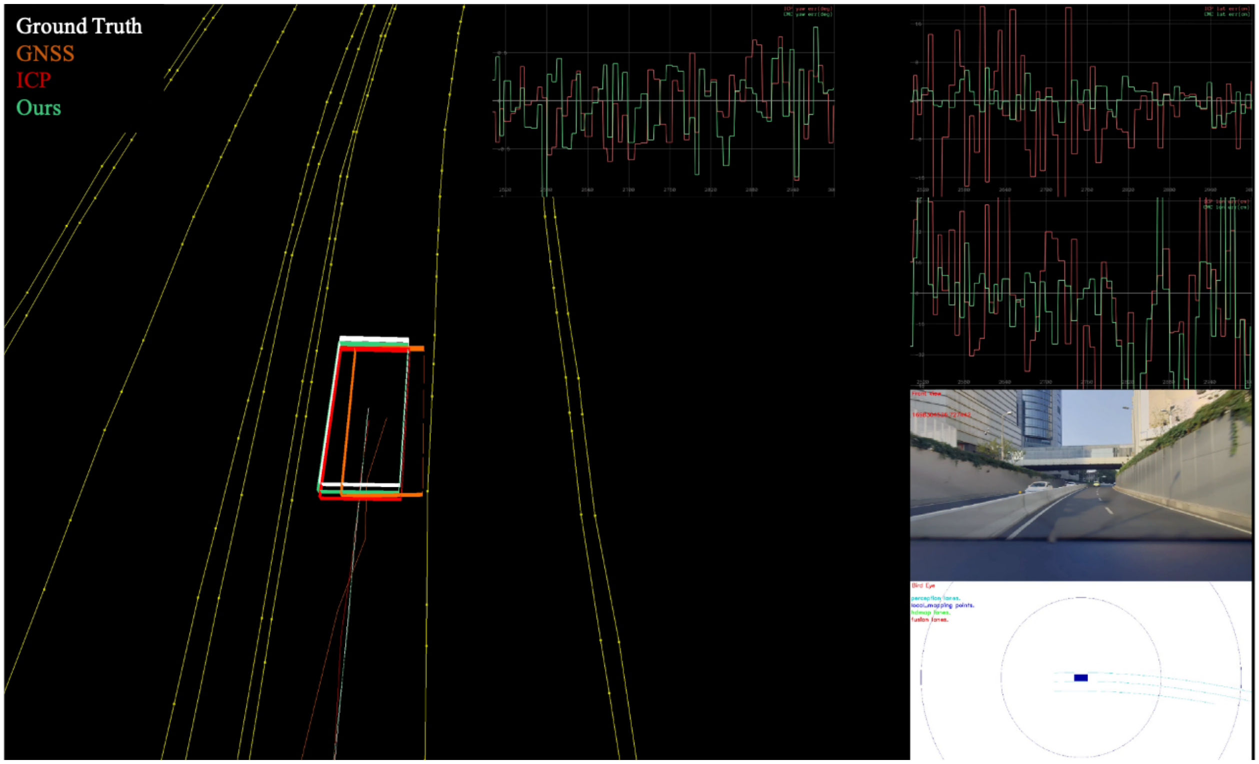

Figure 6. Visualization of Localization Performance in Scenario 2 within the SRoad Dataset.

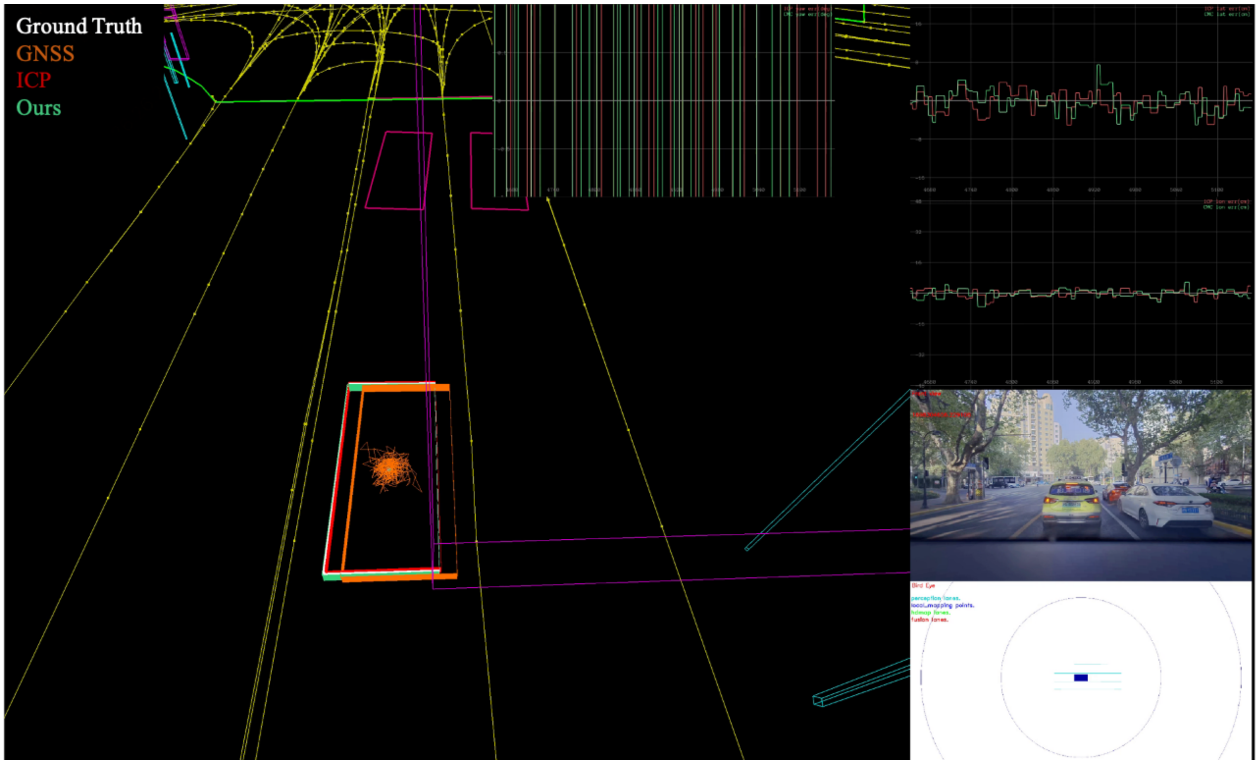

Figure 7. Visualization of Localization Performance in Scenario 3 within the SRoad Dataset.

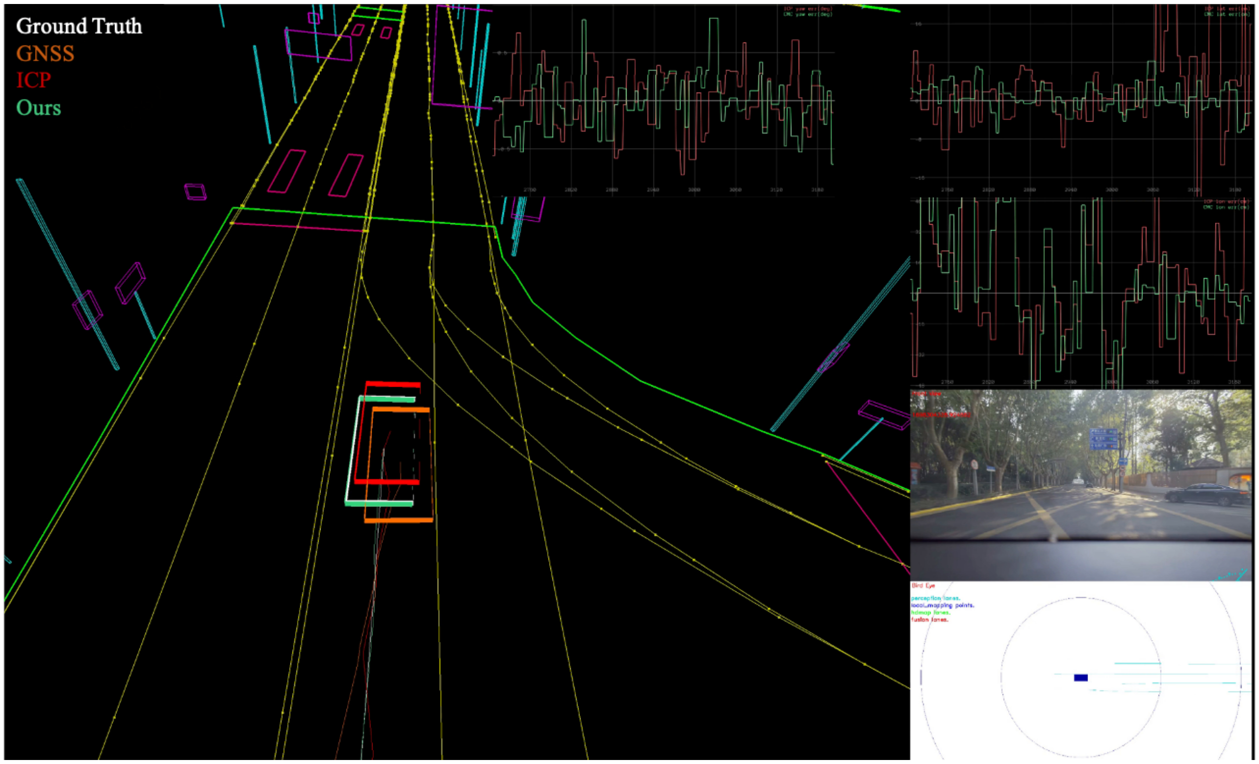

Figure 8. Visualization of Localization Performance in Scenario 4 within the SRoad Dataset.

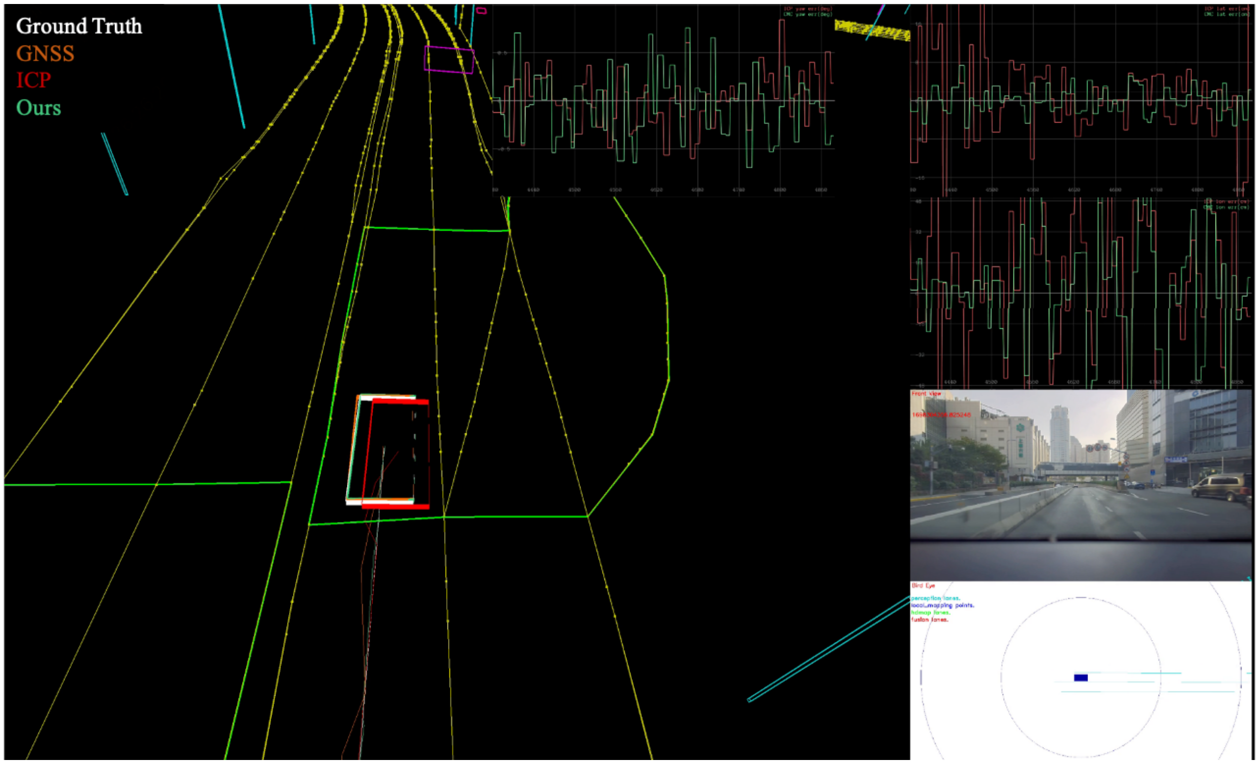

Figure 9. Visualization of Localization Performance in Scenario 5 within the SRoad Dataset.

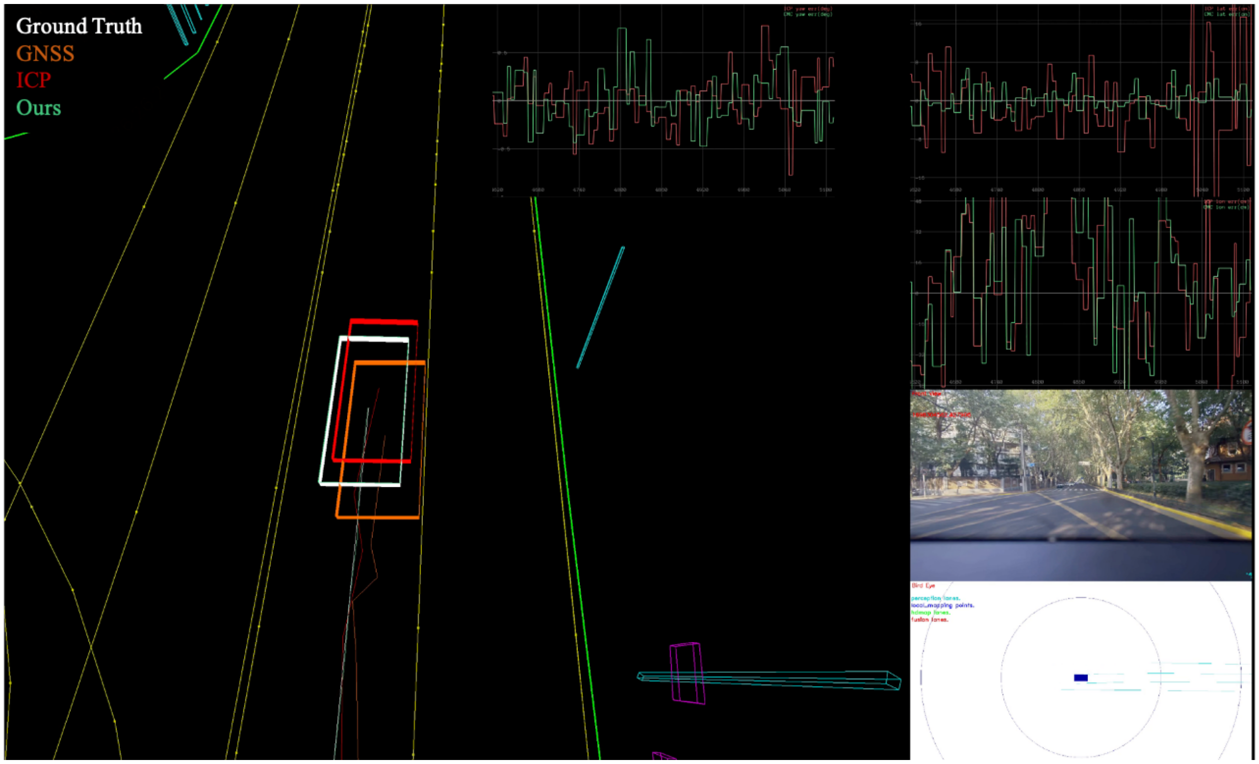

Figure 10. Visualization of Localization Performance in Scenario 6 within the SRoad Dataset.

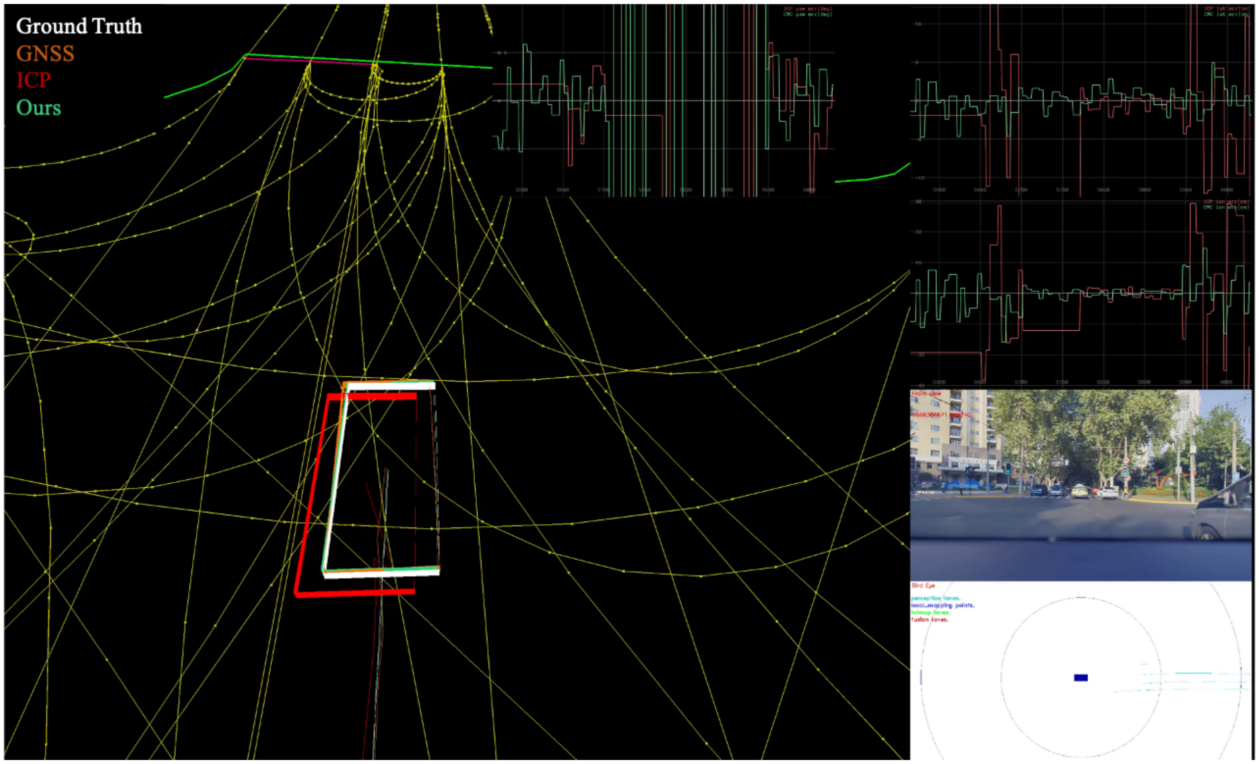

Figure 11. Visualization of Localization Performance in Scenario 7 within the SRoad Dataset.

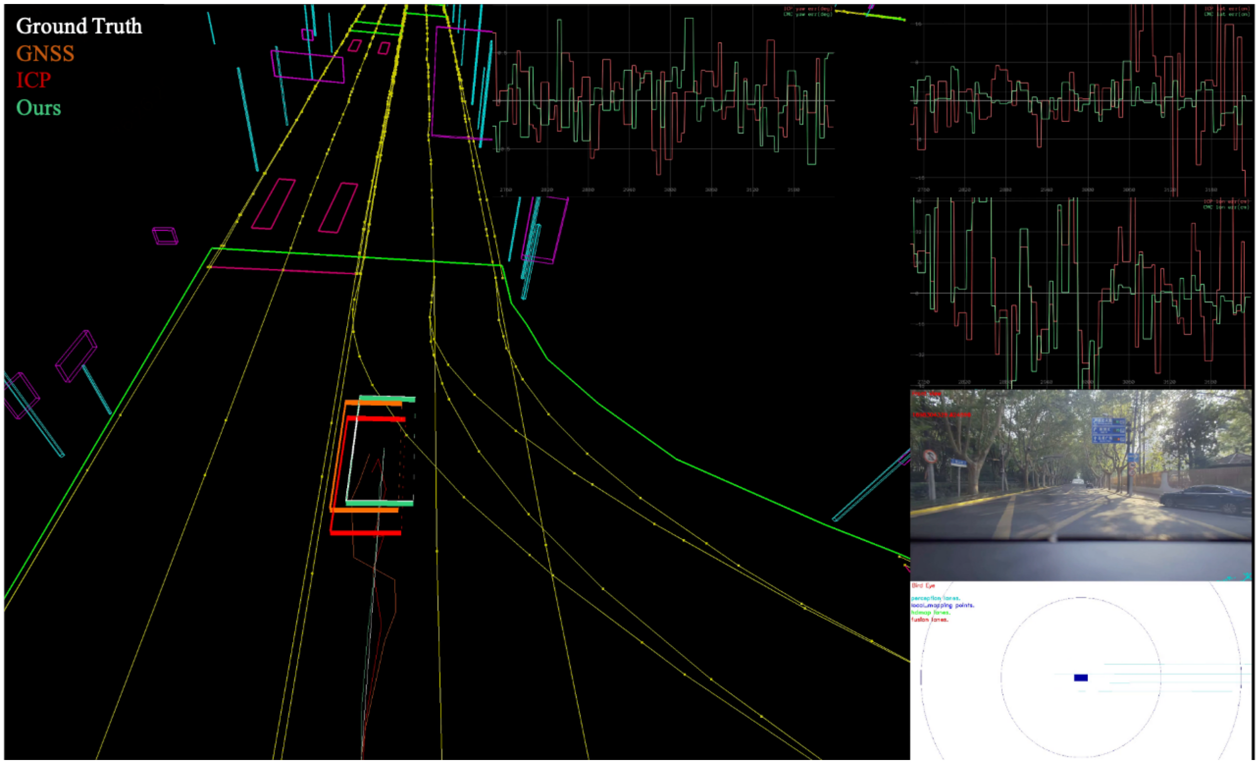

Figure 12. Visualization of Localization Performance in Scenario 8 within the SRoad Dataset.

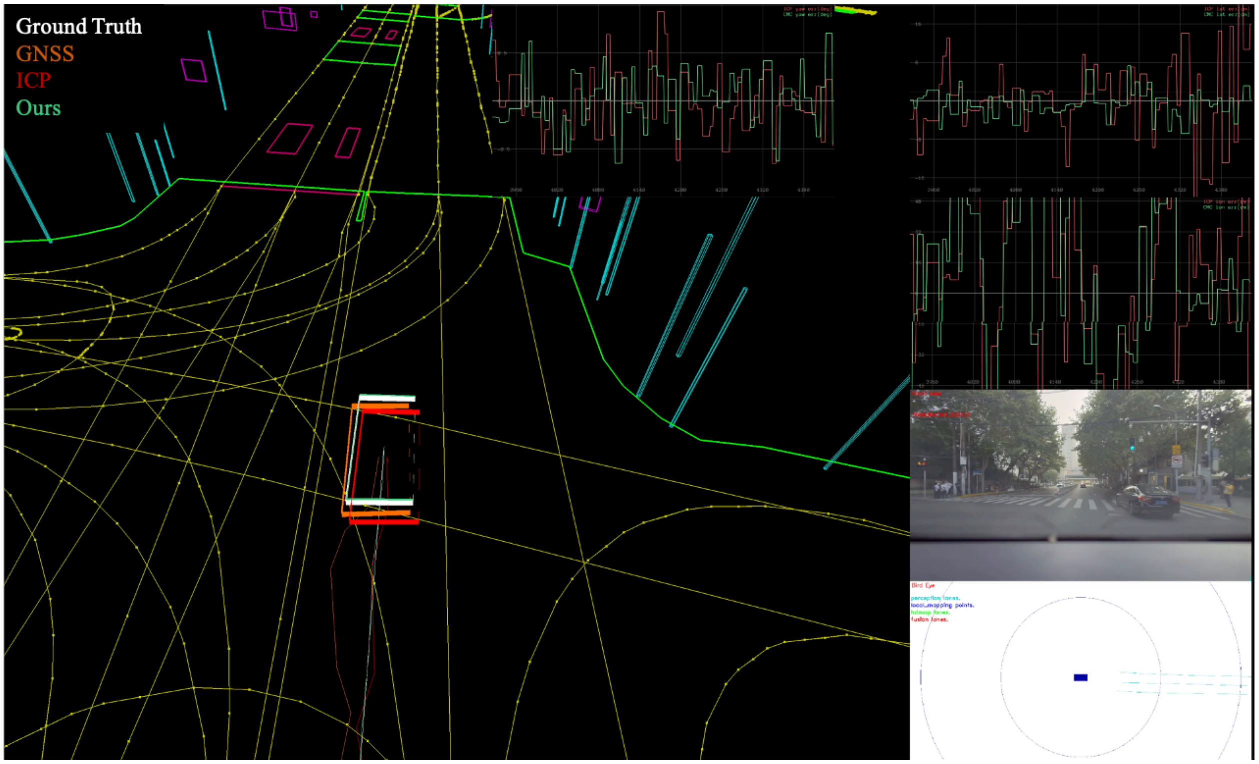

Figure 13. Visualization of Localization Performance in Scenario 9 within the SRoad Dataset.

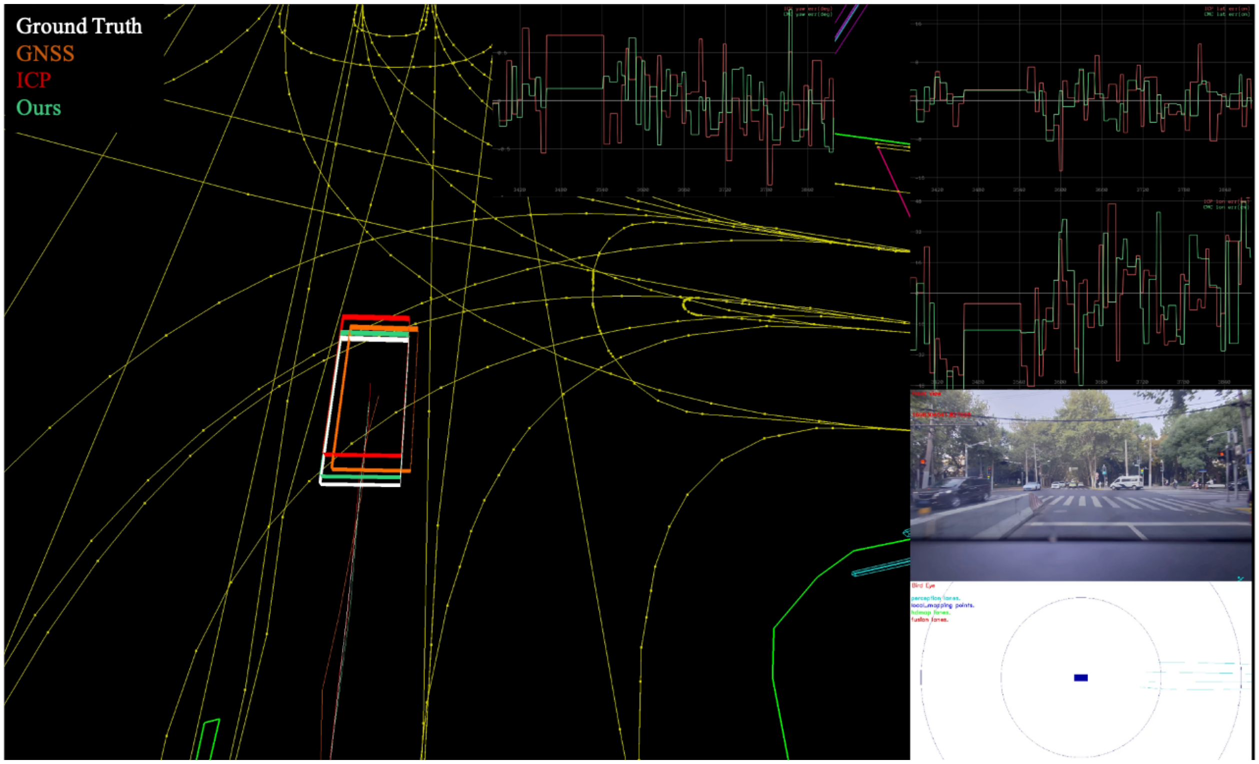

Figure 14. Visualization of Localization Performance in Scenario 10 within the SRoad Dataset.

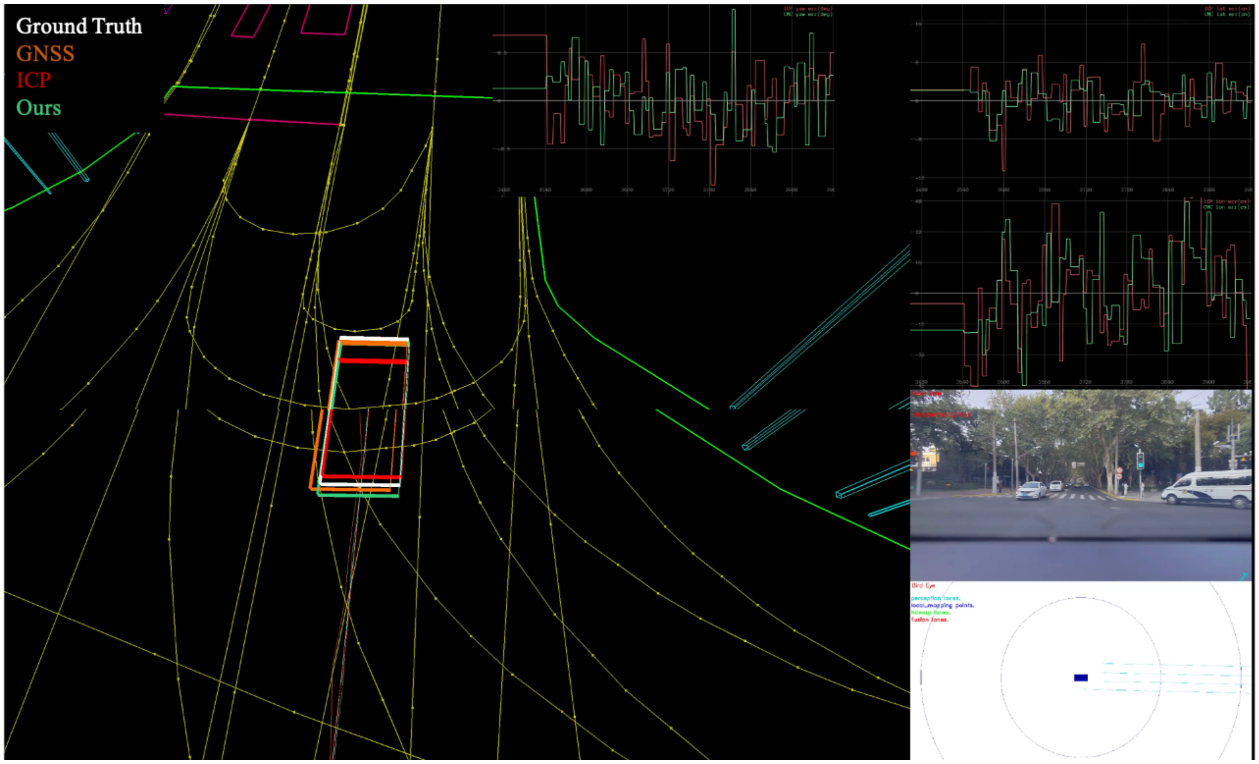

Figure 15. Visualization of Localization Performance in Scenario 11 within the SRoad Dataset.

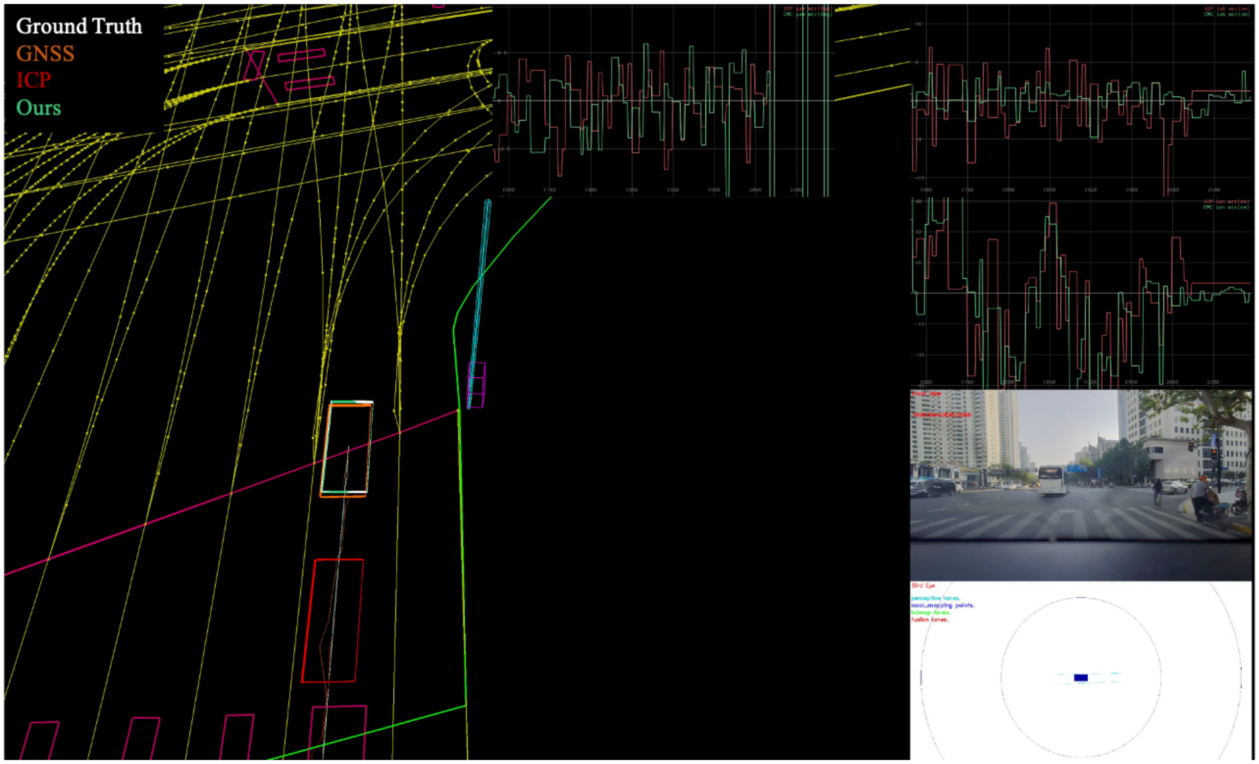

Figure 16. Visualization of Localization Performance in Scenario 12 within the SRoad Dataset.

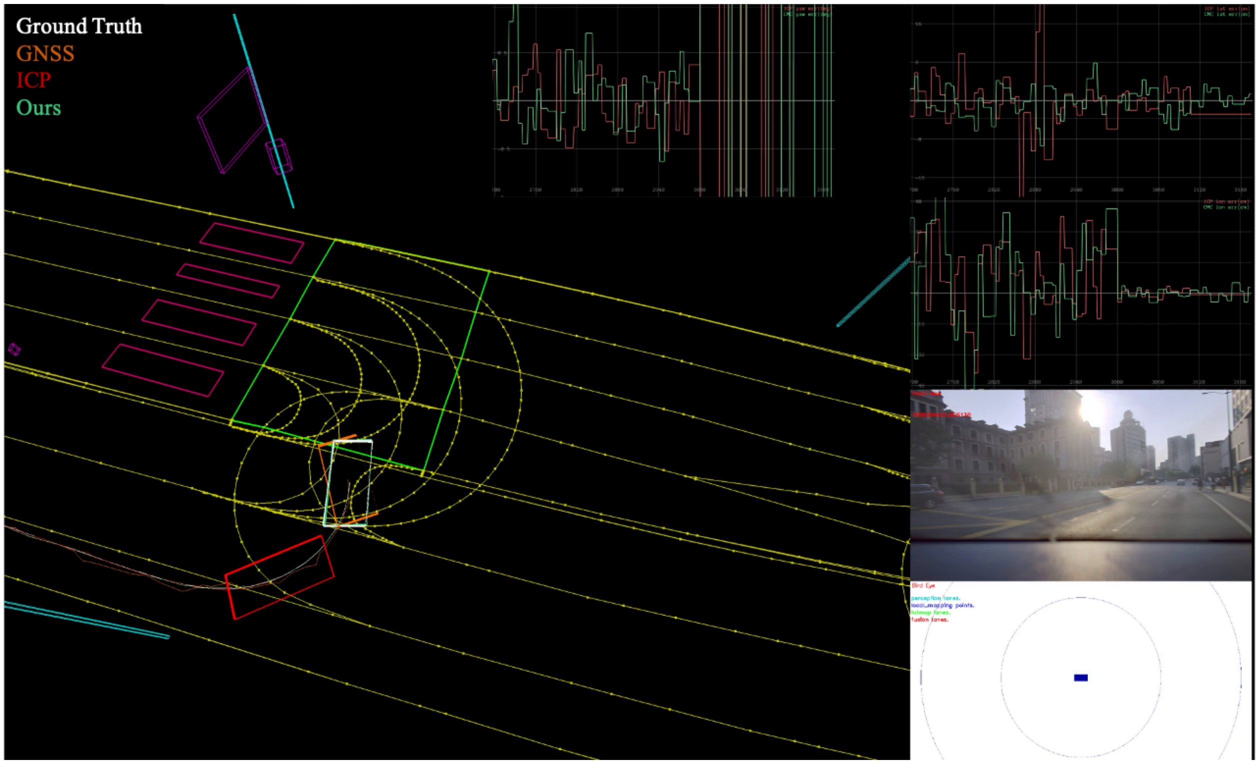

Figure 17. Visualization of Localization Performance in Scenario 13 within the SRoad Dataset.

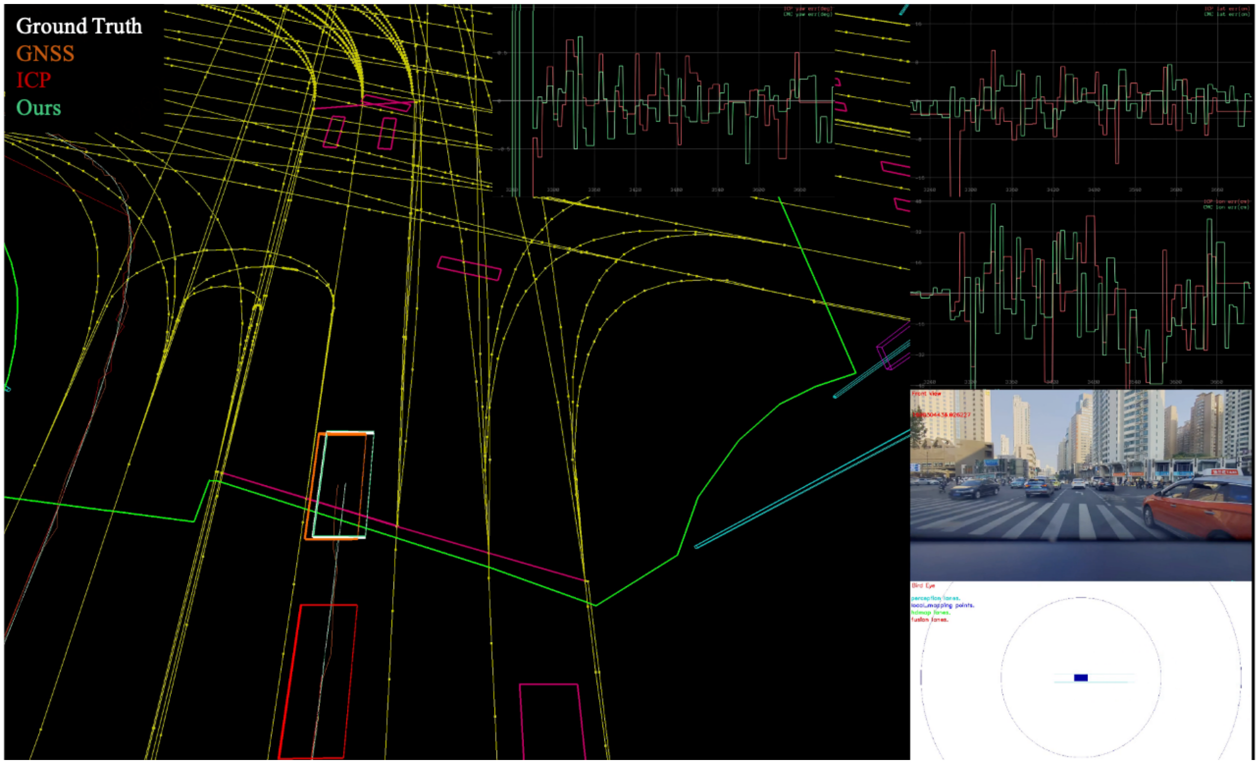

Figure 18. Visualization of Localization Performance in Scenario 14 within the SRoad Dataset.