Supplementary Material:
Beyond Cross-view Image Retrieval:
Highly Accurate Vehicle Localization using Satellite Image

Yujiao Shi and Hongdong Li
The Australian National University, Canberra, Australia
firstname.lastname@anu.edu.au

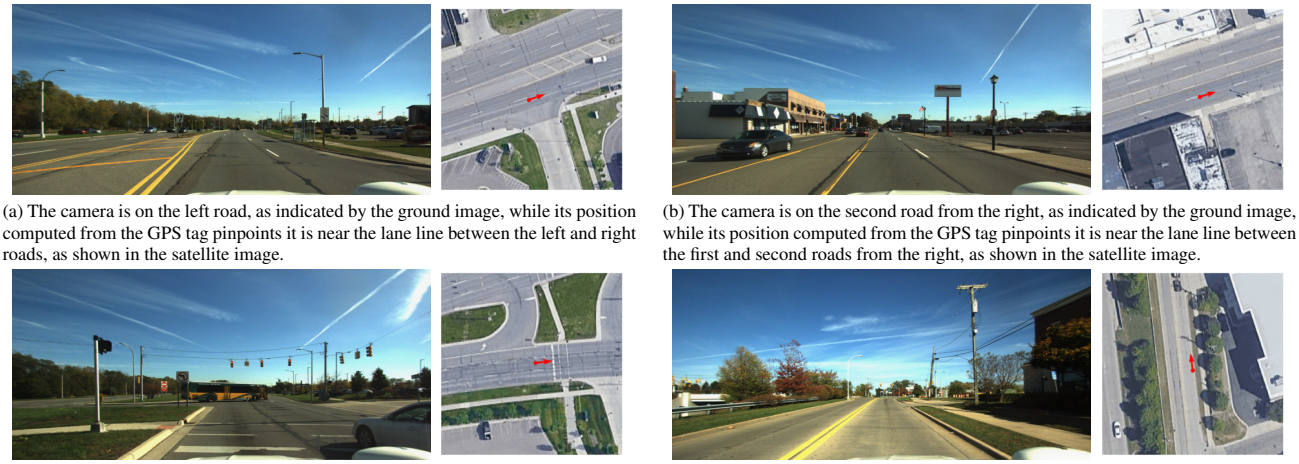
1. Training and Testing Splits of the KITTI and the Ford Multi-AV dataset

Despite both KITTI and the Ford Multi-AV datasets being captured by accurate survey-grade RTK-GPS systems, we have uncovered that their ground-truth GPS tags are sometimes contaminated by considerable noises. This can be seen, for example, by marking up the GPS-reported camera position in the satellite image and visually comparing if the observed ground-level scenes as if seen from the ground plane matches well with the marked position in the satellite image. Fig. 1 and Fig. 2 illustrate some examples from the Ford Dataset, which clearly reveal such mismatches.

We manually filter out those inaccurate ones and construct new subsets for the KITTI and the Ford multi-AV dataset to train and evaluate our new localization method. The training and testing image numbers of the two datasets are presented in Tab. 1 and Tab. 2, respectively.

To validate such a pre-filtering is necessary, we conducted comparisons between "training on the full dataset" and "training on the filtered dataset" on the first two logs of the Ford multi-AV dataset. The results are presented in Tab. 3. They are evaluated on the same test sets for fair comparisons. It can be seen that the pre-filtering strategy significantly boosts the performance, especially for lateral translation optimization.

We provide the performance of our method on the remaining logs (Log3∼Log6) of the Ford multi-AV dataset in Tab. 4, to complement our results in Sec. 6.1 of the main paper.



(a) The camera is on the left road, as indicated by the ground image, while its position computed from the GPS tag pinpoints it is near the lane line between the left and right roads, as shown in the satellite image.

(b) The camera is on the second road from the right, as indicated by the ground image, while its position computed from the GPS tag pinpoints it is near the lane line between the first and second roads from the right, as shown in the satellite image.

(c) The camera is on the left of the main road, as indicated by the ground image, while its position computed from the GPS tag pinpoints it is in the middle of the main road, as shown in the satellite image.

(d) The camera is in the middle of the right road, as indicated by the ground image, while its position computed from the GPS tag pinpoints it is near the right boundary of the road, as shown in the satellite image.

Figure 1. Examples whose GPS tags are *inaccurate*. In the satellite image of each sub-figure, the red point indicates the camera position computed from the GPS tag, and the red arrow marks the camera facing direction. The images are from Log3 of drive 2017-10-26.

Figure 2. Examples whose GPS tags are *accurate*. In the satellite image of each sub-figure, the red point indicates the camera position computed from the GPS tag, and the red arrow marks the camera facing direction. The images are from Log2 of drive 2017-10-26.

Table 1. Training and testing image numbers for the KITTI dataset.

	Training	Test1	Test2
#Image	19.655	3,773	7,542

Table 2. Training and testing splits for the Ford multi-AV dataset. (The training and testing sets of Log3 are from the same drive but different locations.)

		Log1	Log2	Log3	Log4	Log5	Log6
Training	Drive	2017-10-26	2017-10-26	2017-08-04	2017-10-26	2017-08-04	2017-08-04
	#Image	4,000	10,350	1,500	7466	8430	3857
Testing	Drive	2017-08-04	2017-08-04	2017-08-04	2017-08-04	2017-10-26	2017-10-26
	#Image	2,100	3,727	1,500	3,511	3,500	1,000

Table 3. Performance comparison of our method on the first two logs of the Ford multi-AV dataset, when trained on the "Full Dataset" or the "Filtered Dataset".

	Log1									Log2								
	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
Full Dataset	26.67	64.76	**79.76**	5.14	15.48	24.14	28.81	66.14	**81.24**	22.14	58.06	71.18	**5.47**	**16.15**	**25.95**	**9.98**	30.35	49.26
Filtered Dataset	**46.10**	**70.38**	72.90	**5.29**	**16.38**	**26.90**	**44.14**	**72.67**	80.19	**31.20**	**66.46**	**78.27**	4.80	15.27	25.76	9.74	**30.83**	**51.62**

Table 4. Performance of our method on the remaining logs of the Ford multi-AV dataset.

	Lateral			Longitudinal			Azimuth				Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$		$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
Log3	11.40	34.00	58.13	4.47	13.13	22.47	8.93	29.73	48.80	Log4	29.96	66.28	74.88	4.96	15.52	25.92	14.33	43.69	67.45
Log5	15.26	54.60	76.71	6.23	19.89	32.34	17.74	47.60	67.74	Log6	20.20	45.20	59.00	3.90	14.30	24.50	10.80	31.80	52.50

2. Increasing the Grid Sample Density for Image Retrieval-based Methods

In this section, we provide additional experiments to investigate the performance of image retrieval-based methods when increasing the grid sample density in constructing the database. Among the state-of-the-art cross-view image retrieval algorithms, DSM [45] and VIGOR [67] are two of the performers. We therefore only compared ours with these two algorithms. From the results in Tab. 5, we did not observe consistent positive effects when increasing the grid sample density. This might be because, in the fine-grained retrieval-based localization, the database images using a grid of 4×4 are already very similar and hard to discriminate. Thus, increasing the sample density of database images does not help. Fig. 3 presents some examples of the database images sampled using a grid of 4×4.

3. Different Initial Values

In Tab. 6, we show the performance of our method with different pose initialization ranges. The performance increases as the search range decreases. The consumer-level GPS accuracy ranges from 15m to 20m, and the image retrieval methods [44, 54] can make their top-1 retrieved results be within 5m to their ground truth. Since the primary purpose of this paper is

Figure 3. The database images for fine-grained image retrieval using a grid of 4×4. They are very similar and hard to discriminate.

Table 5. Performance of image retrieval-based methods when increasing the grid sample density on the KITTI dataset.

	Grid	Test1									Test2								
		Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
		$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
DSM [45]	4×4	12.00	35.29	53.67	4.33	12.48	21.43	3.52	13.33	23.67	8.45	24.85	37.64	3.94	12.24	21.41	2.23	7.67	13.42
	5×5	11.69	33.34	50.25	4.51	13.68	21.55	3.66	13.65	24.49	11.44	33.16	50.76	4.11	12.13	20.35	3.20	13.35	23.67
	6×6	12.72	34.35	50.15	4.53	12.70	21.89	3.45	13.65	24.44	12.25	34.31	51.83	4.04	12.49	21.13	3.37	13.55	23.77
	7×7	12.80	35.38	50.41	4.93	13.60	22.55	3.60	13.91	25.10	12.42	34.91	51.72	3.99	12.56	21.49	3.31	13.14	23.38
VIGOR [67]	4×4	20.33	52.48	70.43	6.19	16.05	25.76	-	-	-	20.87	54.87	75.64	**5.98**	**16.88**	**27.23**	-	-	-
	5×5	18.98	48.85	70.34	4.59	13.89	22.77	-	-	-	16.83	48.38	71.15	4.08	12.32	20.91	-	-	-
	6×6	17.84	48.98	70.39	5.17	14.58	24.07	-	-	-	17.54	48.46	71.40	4.46	13.56	22.01	-	-	-
	7×7	18.50	49.06	70.55	4.90	14.15	23.43	-	-	-	17.37	48.48	71.68	4.36	13.71	22.29	-	-	-
Ours	-	**35.54**	**70.77**	**80.36**	**5.22**	**15.88**	**26.13**	**19.64**	**51.76**	**71.72**	**27.82**	**59.79**	**72.89**	5.75	16.36	26.48	**18.42**	**49.72**	**71.00**

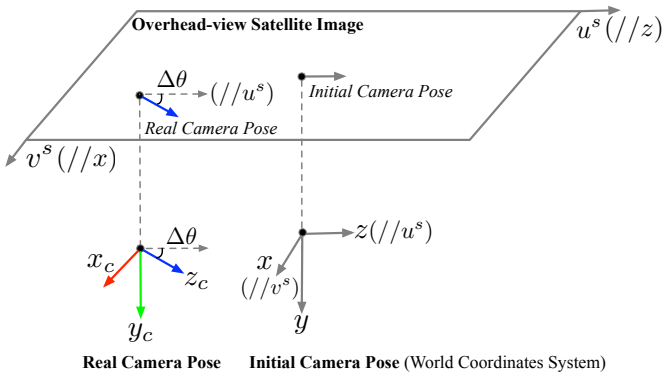


Figure 4. Coordinates illustration. Note that this is only for illustration purpose. The coordinates used in our codes are slightly different with this one.

to study whether we can refine an initial coarse estimate by cross-view matching, we set our search region as 40m×40m in the paper.

4. Additional Comparisons

Ours w/o Long. We investigate whether the loss item on longitudinal pose estimation can be removed, denoted as "Ours w/o Long". As shown in the first row of Tab. 7, this results in a negative effect, indicating that the longitudinal pose constraints contribute to learning discriminative features, although the ambiguity along this direction is high.

Different iteration strategies. In our framework, the LM optimization is first applied to the multi-level features from coarse to fine (C2F), and then the C2F update is executed iteratively. Here, we study the performance of the LM optimization when it is first applied to the coarsest feature level until the maximum iteration and then propagates to finer levels, denoted as "C2F Global". The results are presented in the second row of Tab. 7. Compared to C2F Global, our update strategy guarantees fine-tuning around more possible solutions and thus is more likely to find the global optimum.

Table 6. Performance comparison with different search regions on the KITTI dataset.

Search Region	Test1									Test2								
	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
40m×40m	35.54	70.77	80.36	5.22	15.88	26.13	19.64	51.76	71.72	27.82	59.79	72.89	5.75	16.36	26.48	18.42	49.72	71.00
20m×20m	44.66	73.92	81.18	12.06	35.62	54.73	25.31	57.41	74.48	34.17	72.30	81.15	11.56	35.08	53.77	11.40	48.18	65.80
10m×10m	64.86	92.23	96.98	29.08	69.49	88.66	36.92	73.95	86.88	55.98	90.84	96.43	25.97	66.96	88.12	31.36	69.46	84.50

Table 7. Additional ablation study results of our method on the KITTI dataset.

	Test1									Test2								
	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
Ours w/o Long	25.63	56.72	69.55	**5.99**	**16.06**	**26.85**	13.84	39.01	59.98	20.50	52.52	67.57	5.32	15.16	25.23	12.90	36.79	57.73
C2F Global	23.32	50.60	61.25	5.27	15.88	26.05	11.87	33.66	54.86	20.43	45.86	58.51	5.25	15.82	26.16	11.65	33.65	54.02
Ours	**35.54**	**70.77**	**80.36**	5.22	15.88	26.13	**19.64**	**51.76**	**71.72**	**27.82**	**59.79**	**72.89**	**5.75**	**16.36**	**26.48**	**18.42**	**49.72**	**71.00**

5. Coordinates Illustration and Pose Parameterization

We set the world coordinates system to the initial camera pose estimate, as shown in Fig. 4. For illustration brevity, we pre-align the satellite image to make its center correspond to the initial camera position and its u direction parallel to the initial camera facing direction. Here, both z and z_c denote the camera facing direction.

Denote Δx is the lateral translation, Δz is the longitudinal translation, and θ is the azimuth angle. The query ground camera pose in Eq. (2) and Eq. (4) in the main paper is parameterized as

$$\mathbf{R} = \begin{bmatrix} -\sin\theta & 0 & \cos\theta \\ \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} \Delta x \\ 0 \\ \Delta z \end{bmatrix}. \tag{1}$$

6. Broader Impact

This paper has introduced a new technique for high-accuracy vehicle/camera localization. This technique can provide accurate camera position estimation even in a GPS-denied environment. The position of a vehicle or camera of a user is often considered sensitive or private information. The proposed technique may be abused or misused, causing privacy violations. We advocate careful data protection and model management to mitigate the risk.