Supplementary Material: Boosting 3-DoF Ground-to-Satellite Camera Localization Accuracy via Geometry-Guided Cross-View Transformer

Yujiao Shi[1], Fei Wu[1], Akhil Perincherry[2], Ankit Vora[2] and Hongdong Li[1]
[1]The Australian National University [2]Ford Motor Company
yujiao.shi@anu.edu.au

1. Performance on Log3-6 in Ford Multi-AV

We present the performance of deep LM [29], CVML [43] and our method on remaining logs in the cross-view Ford multi-AV dataset with aligned-orientation in Tab. 1. It can be seen that our method achieves significantly better performance in most cases. Since CVML cannot estimate orientation, we only compare with deep LM on joint location and orientation estimation. The results are provided in Tab. 2. Our method achieves consistently better performance.

Table 1. Performance comparison on remaining logs of Ford Multi-AV with aligned-orientation.

	Lateral		Longitudinal		Lateral		Longitudinal		Lateral		Longitudinal		Lateral		Longitudinal	
	$d=1$	$d=3$	$d=1$	$d=3$	$d=1$	$d=3$	$d=1$	$d=3$	$d=1$	$d=3$	$d=1$	$d=3$	$d=1$	$d=3$	$d=1$	$d=3$
	Ford - Log3				Ford - Log4				Ford - Log5				Ford - Log6			
LM [29]	20.40	43.27	5.13	14.80	41.81	79.86	6.61	19.14	23.86	59.11	4.37	18.86	15.6	46.3	5.1	14.9
CVML [43]	10.07	42.27	4.33	15.4	44.94	80.23	13.05	31.36	16.77	49.31	5.57	17.06	**29.6**	66.2	4.8	16.2
Ours	**36.67**	**47.47**	**6.67**	**19.20**	**88.38**	**99.12**	**34.72**	**63.23**	**26.69**	**71.86**	**13.77**	**29.14**	27.3	**75.9**	**11.9**	**24.1**

Table 2. Performance comparison on remaining logs of Ford Multi-AV with 20° orientation noise.

	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
	Ford - Log3									Ford - Log4								
LM [29]	11.40	34.00	58.13	4.47	13.13	22.47	8.93	29.73	48.80	29.96	66.28	74.88	4.96	15.52	25.92	14.33	43.69	67.45
Ours	**29.40**	**57.87**	**63.00**	**6.07**	**17.53**	**25.00**	**26.67**	**61.93**	**90.20**	**79.58**	**97.78**	**98.72**	**15.98**	**39.50**	**53.77**	**64.17**	**97.04**	**99.49**
	Ford - Log5									Ford - Log6								
LM [29]	15.26	54.60	76.71	6.23	19.89	32.34	17.74	47.60	67.74	20.20	45.20	59.00	3.90	14.30	24.50	10.80	31.80	52.50
Ours	**22.66**	**70.14**	**90.71**	**15.43**	**31.80**	**38.74**	**63.71**	**88.83**	**93.71**	**30.80**	**68.70**	**78.90**	**9.70**	**23.70**	**29.10**	**55.10**	**93.00**	**98.60**

2. Why Ground-to-Satellite Synthesis

The satellite-to-ground (S2G) synthesis has been demonstrated to be superior to ground-to-satellite (G2S) synthesis in joint location and orientation optimization by Shi and Li [29]. This section demonstrates that G2S synthesis is more essential than S2G synthesis for pure orientation optimization, especially when the orientation ambiguity is significant.

Particularly, S2G synthesis leads to significant information loss, as a ground-view image only corresponds to a portion of the satellite image. Consider two ground-view images at the same location. Only when their orientation difference is slight can their overlap be sufficient to predict their relative orientation. Their overlap will be significantly reduced when a more considerable orientation difference occurs, making orientation estimation intractable. In contrast, G2S synthesis reserves all information from the ground image, and its comparison counterpart (the satellite image) contains the entire scene content. Hence, a larger possibility of correct orientation estimation can be guaranteed.

Table 3. Performance comparison on KITTI with increasing orientation noise.

Rotation Noise		Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
		$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
		Test1									Test2								
20°	S2G	70.02	94.96	97.80	16.22	38.14	48.50	87.57	99.95	**100.00**	56.72	**89.63**	**93.69**	12.76	30.81	41.63	87.52	99.89	**100.00**
	G2S (Ours)	**76.44**	**96.34**	**98.89**	**23.54**	**50.57**	**62.18**	**99.10**	**100.00**	**100.00**	**57.72**	86.77	91.16	**14.15**	**34.59**	**45.00**	**98.98**	**100.00**	**100.00**
80°	S2G	54.17	88.07	94.17	13.46	33.08	43.94	19.56	51.23	70.13	46.90	84.65	92.15	11.92	28.52	38.86	21.82	54.83	74.18
	G2S (Ours)	**70.21**	**95.47**	**98.28**	**22.29**	**48.90**	**59.50**	**53.27**	**93.98**	**98.99**	**56.97**	**87.72**	**92.35**	**15.17**	**35.39**	**47.02**	**58.68**	**95.92**	**99.15**

We compare G2S and S2G synthesis for orientation estimation under the proposed pipeline. The results are provided in Tab. 3. Note that the location estimation in this comparison is kept the same as in our proposed method. It can be seen that the rotation estimation accuracy by G2S synthesis is always better than that of S2G synthesis. When the rotation noise becomes significant, the performance discrepancy becomes larger. Benefiting from the high rotation estimation performance by G2S synthesis, the location estimation performance under the same pipeline is also better.

The other advantage of G2S over S2G synthesis is its superior efficiency in location estimation when the orientation has already been estimated. G2S synthesis retains all the information from the ground view, and the synthesized features for different translations are the same (with only a translation difference between the synthesized feature maps). As a result, the projection only needs to be conducted once. Then, a fast spatial correlation using Pytorch/TensorFlow built-in layer can be used to implement the dense search, enabling high computational efficiency.

In contrast, S2G projected features for different translations are different, and these differences are crucial for translation estimation. Therefore, S2G projection needs to be conducted as many times as the candidate translation poses when implementing dense search, leading to unaffordable computation and memory consumption. This is also why we cannot compare the performance of G2S and S2G synthesis on location estimation using the dense search mechanism.

3. Additional Model Analysis

Here, we conduct further ablation studies to demonstrate the effectiveness and necessity of each component in the proposed method. Starting from the deep LM [29] for joint rotation and translation optimization, we first replace the deep LM optimizer with the proposed Neural pose Optimizer (Neural Opt.). The first two rows of Tab. 4 indicate that the proposed neural pose optimizer achieves significantly better performance in rotation estimation but worse performance in translation estimation. This demonstrates our intuition that the translation on the neural optimizer input will be absorbed by the deep high-level features inside the optimizer, thus predicting inaccurate translation estimates.

Next, we construct a decoupling framework where the neural pose optimizer is employed for rotation estimation while the deep LM optimizer is adopted for translation optimization. The third row of Tab. 4 shows the neural optimizer's powerful rotation estimation performance is mainly inherited in the decoupling framework, benefiting from which the deep LM's lateral translation optimization performance improves. This demonstrates that it is beneficial to develop different rotation and translation estimation strategies to maximize their performance.

Then, we replace the deep LM optimizer in the decoupling framework with the proposed Dense Search (DS) mechanism (spatial correlation between synthesized overhead view feature map and observed satellite image feature map). From the fourth row of Tab. 4, it can be seen the performance on both rotation and translation increases, indicating the superiority of the proposed dense search mechanism over deep LM optimization. The last two rows of Tab. 4 present the performance when employing the proposed geometry-guided cross-view transformer for overhead view feature synthesis and including the uncertainty map in the dense search process, respectively. Both of them contribute to better performance.

Furthermore, we conduct experiments to demonstrate whether it is necessary to apply supervision on the estimated translation by the neural optimizer. Tab. 5 shows the results. It can be seen that the neural optimizer's rotation estimation performance is poor when its translation predictions are not supervised. This indicates that additional supervision on the neural optimizer translation predictions provides more clues for the neural optimizer weight learning. This is especially important because the feature extractors and the neural optimizer are randomly initialized. Providing more supervision constraints the network learning freedom.

Table 4. Additional ablation study results of our method on KITTI.

Overhead-view Feature Synthesis	Optimization Scheme	Uncertainty	Test1						Test2					
			Lateral		Longitudinal		Azimuth		Lateral		Longitudinal		Azimuth	
			$d=1$	$d=3$	$d=1$	$d=3$	$\theta=1$	$\theta=3$	$d=1$	$d=3$	$d=1$	$d=3$	$\theta=1$	$\theta=3$
Geometry projection	LM ($\mathbf{R}$ & $\mathbf{t}$)	N/A	27.72	59.98	5.75	16.8	18.13	48.77	27.82	59.79	5.75	16.36	18.42	49.72
	Neural Opt. ($\mathbf{R}$ & $\mathbf{t}$)	N/A	4.90	15.00	4.88	15.66	97.75	100.00	5.09	15.39	5.32	15.87	97.75	100.00
	Neural Opt. ($\mathbf{R}$) + LM ($\mathbf{t}$)	N/A	41.82	76.28	5.27	15.74	87.01	100.00	28.53	64.82	5.42	16.11	87.79	100.00
	Neural Opt. ($\mathbf{R}$) + DS ($\mathbf{t}$)	No	64.83	92.15	14.47	33.87	99.92	100.00	51.70	79.79	9.45	24.11	99.80	100.00
Geometry-guided Cross-view Transformer	Neural Opt. ($\mathbf{R}$) + DS ($\mathbf{t}$)	No	70.29	94.30	18.29	40.39	84.44	99.84	54.53	85.44	12.50	29.65	84.16	99.80
	Neural Opt. ($\mathbf{R}$) + DS ($\mathbf{t}$)	Yes	**76.44**	**96.34**	**23.54**	**50.57**	**99.10**	**100.00**	**57.72**	**86.77**	**14.15**	**34.59**	**98.98**	**100.00**

Table 5. Performance of our method with or without translation supervision applied on the proposed neural pose optimizer on KITTI.

Translation supervision on neural pose optimizer	Test1									Test2								
	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
No	50.44	20.59	93.90	46.75	98.41	57.86	10.52	30.19	50.52	41.16	13.23	82.91	30.93	90.73	41.59	10.10	29.94	50.61
Yes	76.44	96.34	98.89	23.54	50.57	62.18	99.10	100.00	100.00	56.97	87.72	92.35	15.17	35.39	47.02	58.68	95.92	99.15

4. Number of Iterations of the Proposed Neural Pose Optimizer

In this section, we study the choice of iteration number for the proposed neural optimizer. Results are presented in Tab. 6. It can be seen that using two iterations contributes to a better performance than using one iteration. However, the performance becomes robust/similar when increasing the iteration number further. Thus, we use the iteration number as two in our proposed framework.

Table 6. Performance comparison on KITTI with different iteration numbers of the proposed neural optimizer.

No. Iterations	Test1									Test2								
	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
1	75.19	95.68	97.93	21.23	47.15	57.96	99.87	100.00	100.00	53.94	84.23	89.72	13.52	33.90	45.35	99.80	100.00	100.00
2	76.44	96.34	98.89	23.54	50.57	62.18	99.10	100.00	100.00	57.72	86.77	91.16	14.15	34.59	45.00	98.98	100.00	100.00
3	75.85	96.16	98.54	23.32	50.68	61.17	80.52	99.89	100.00	58.04	86.58	90.77	13.80	33.43	44.78	77.53	100.00	100.00
4	76.60	96.50	98.75	23.32	50.41	60.69	79.91	99.95	100.00	58.21	87.88	92.12	15.02	35.93	46.91	77.34	100.00	100.00
5	75.30	95.73	98.33	22.71	50.78	61.17	81.13	99.97	100.00	58.98	88.50	92.54	15.02	35.85	46.80	78.23	99.88	100.00

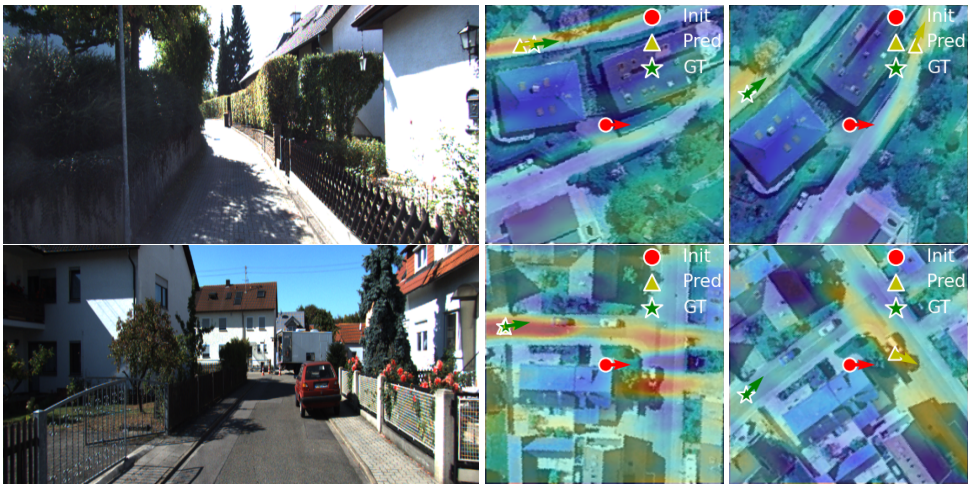
5. Different Initial Values

Orientation. Below, we increase the rotation noise and compare the performance of our method with deep LM. The location search range follows the official setting [29]: within a 40m $\times$ 40m search space. The rotation noise is set to $20°$, $40°$, and $80°$. From the results in Tab. 7, it can be seen that deep LM almost fails on rotation estimation when the rotation noise increases to $80°$, with only 9% of the images whose estimated rotation is within $3°$ to its ground truth values. In contrast, our method makes over 90% of the estimated rotation within $3°$ of their GT values. Since the rotation ambiguity has been significantly reduced, the translation estimation performance of our method is robust.

Tab. 8 provides the quantitative evaluation of our method when increasing the rotation noise to $180°$. It can be seen that the rotation estimation performance drops when the rotation noise increases to a large number, which also affects the translation estimation performance. Fig. 1 shows two examples. The query images in Fig. 1 tell that the cameras are facing towards the road. When the rotation prior is relatively accurate, we can quickly determine the camera's orientation by comparing the rotation angles between the initialized orientation and the road directions. Then, the camera location will also be estimated on the right road, as shown in the middle column of Fig. 1. However, when the rotation ambiguity is considerable, there is a large probability of the query image being matched to a wrong part of the satellite image, resulting incorrect rotation and translation estimates, as shown in the right column of Fig. 1. In practice, we suggest first restricting the rotation ambiguity of the query cameras to an acceptable range(*e.g.*, within $80°$) and then adopting ground-to-satellite image matching to refine this pose.

Table 7. Performance comparison on KITTI with increasing orientation noise.

Rotation Noise		Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
		$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
		Test1									Test2								
20°	LM [29]	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
	Ours	**76.44**	**96.34**	**98.89**	**23.54**	**50.57**	**62.18**	**99.10**	**100.00**	**100.00**	**57.72**	**86.77**	**91.16**	**14.15**	**34.59**	**45.00**	**98.98**	**100.00**	**100.00**
40°	LM [29]	32.02	68.43	79.25	5.46	16.57	27.46	13.70	36.47	52.74	26.58	62.94	74.93	5.34	16.12	26.47	10.82	32.38	48.69
	Ours	**70.42**	**94.75**	**97.43**	**20.49**	**48.50**	**58.49**	**69.12**	**99.68**	**99.97**	**55.52**	**86.20**	**91.10**	**13.66**	**33.74**	**44.95**	**76.45**	**99.59**	**99.99**
80°	LM [29]	26.95	62.39	78.40	5.14	15.69	26.27	3.10	8.88	15.00	22.43	54.63	71.03	5.17	15.78	25.97	3.05	8.50	14.25
	Ours	**70.21**	**95.47**	**98.28**	**22.29**	**48.90**	**59.50**	**53.27**	**93.98**	**98.99**	**56.97**	**87.72**	**92.35**	**15.17**	**35.39**	**47.02**	**58.68**	**95.92**	**99.15**



Query Image 20° Rotation Ambiguity 100° Rotation Ambiguity

Figure 1. Localization results of our method when rotation ambiguity is different. The satellite images in each row are from the same place with different rotations to mimic the different rotation noises of the ground camera. When the orientation noise is large (100°), the rotation estimation ambiguity becomes significant because the limited scene content captured by a query image can be matched to the different parts on a satellite image.

Table 8. Performance of our method on KITTI with different rotation noises.

Rotation Ambiguity	Test1									Test2								
	Lateral			Longitudinal			Azimuth			Lateral			Longitudinal			Azimuth		
	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$	$d=1$	$d=3$	$d=5$	$d=1$	$d=3$	$d=5$	$\theta=1$	$\theta=3$	$\theta=5$
20°	76.44	96.34	98.89	23.54	50.57	62.18	99.10	100.00	100.00	57.72	86.77	91.16	14.15	34.59	45.00	98.98	100.00	100.00
40°	70.42	94.75	97.43	20.49	48.50	58.49	69.12	99.68	99.97	55.52	86.20	91.10	13.66	33.74	44.95	76.45	99.59	99.99
60°	72.78	95.02	98.07	21.60	47.79	58.20	60.75	97.80	99.84	54.85	85.76	90.86	14.12	33.36	45.12	67.78	98.36	99.76
80°	70.21	95.47	98.28	22.29	48.90	59.50	53.27	93.98	98.99	56.97	87.72	92.35	15.17	35.39	47.02	58.68	95.92	99.15
100°	50.78	83.28	91.44	16.94	39.04	49.83	20.67	51.55	67.00	40.82	75.10	85.32	13.59	32.37	42.91	18.97	50.98	68.39
120°	37.26	71.38	85.13	14.92	34.96	45.51	13.07	34.53	46.57	30.69	63.96	77.49	11.31	28.52	38.44	11.48	30.67	43.46
140°	29.42	62.10	77.52	10.39	25.87	35.70	7.55	20.70	31.25	22.83	52.74	69.24	8.51	21.04	29.75	6.93	18.56	26.72
160°	24.83	56.19	73.63	10.52	26.45	36.05	5.78	17.52	25.68	19.08	46.43	62.19	8.15	20.70	30.22	4.53	13.46	21.81
180°	16.94	45.77	63.80	7.77	20.81	29.92	3.47	9.70	16.09	14.86	39.54	57.44	7.11	18.96	27.46	2.56	8.17	14.00

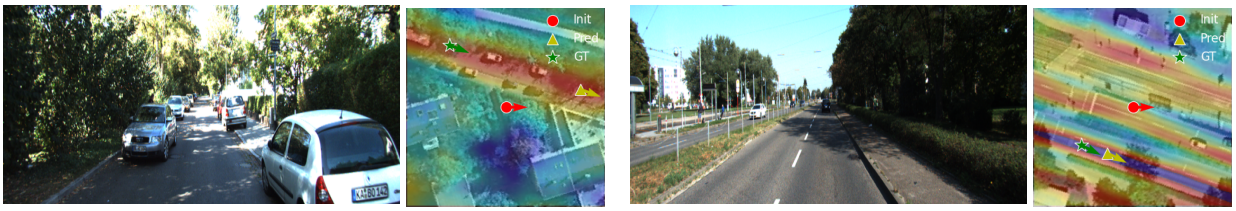
Location. Next, we investigate the performance of our method with different location initialization ranges. Tab. 9 presents the comparison between our approach and deep LM [29]. It can be seen that both methods achieve better performance when the location initialization range is smaller. The performance gap between deep LM and our method increases as the location initialization range increases. This indicates our method is more robust to the location initialization ranges/errors than LM because of the proposed dense search strategy for location estimation. Tab. 10 provides the results of our method when we keep increasing the location search range. The performance decreases gradually as the location search range increases. But even with a search range of 100m $\times$ 100m, our method still outperforms LM when its location search range is 20m $\times$ 20m.

Table 9. Performance comparison between LM[29] and our method with different location priors and 20° orientation noise.

Initialization Range	Methods	Test1 Lateral $d=1$	Test1 Lateral $d=3$	Test1 Lateral $d=5$	Test1 Longitudinal $d=1$	Test1 Longitudinal $d=3$	Test1 Longitudinal $d=5$	Test1 Azimuth $\theta=1$	Test1 Azimuth $\theta=3$	Test1 Azimuth $\theta=5$	Test2 Lateral $d=1$	Test2 Lateral $d=3$	Test2 Lateral $d=5$	Test2 Longitudinal $d=1$	Test2 Longitudinal $d=3$	Test2 Longitudinal $d=5$	Test2 Azimuth $\theta=1$	Test2 Azimuth $\theta=3$	Test2 Azimuth $\theta=5$
10m x 10m	LM [29]	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
	Ours	**89.37**	**98.67**	**99.81**	**26.72**	**56.03**	**72.36**	**98.20**	**100.00**	**100.00**	**62.16**	**20.49**	**89.87**	**47.97**	**96.66**	**66.12**	**97.44**	**99.95**	**100.00**
20m x 20m	LM [29]	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
	Ours	**85.85**	**98.46**	**99.55**	**23.27**	**46.99**	**58.39**	**98.89**	**99.97**	**100.00**	**60.01**	**14.69**	**87.96**	**35.64**	**92.97**	**48.46**	**99.42**	**100.00**	**100.00**
40m x 40m	LM [29]	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
	Ours	**76.44**	**96.34**	**98.89**	**23.54**	**50.57**	**62.18**	**99.10**	**100.00**	**100.00**	**56.97**	**87.72**	**92.35**	**15.17**	**35.39**	**47.02**	**58.68**	**95.92**	**99.15**

Table 10. Performance of our method with different location priors and 20° orientation noise.

Search Region	Test1 Lateral $d=1$	Test1 Lateral $d=3$	Test1 Lateral $d=5$	Test1 Longitudinal $d=1$	Test1 Longitudinal $d=3$	Test1 Longitudinal $d=5$	Test1 Azimuth $\theta=1$	Test1 Azimuth $\theta=3$	Test1 Azimuth $\theta=5$	Test2 Lateral $d=1$	Test2 Lateral $d=3$	Test2 Lateral $d=5$	Test2 Longitudinal $d=1$	Test2 Longitudinal $d=3$	Test2 Longitudinal $d=5$	Test2 Azimuth $\theta=1$	Test2 Azimuth $\theta=3$	Test2 Azimuth $\theta=5$
10m x 10m	89.37	98.67	99.81	26.72	56.03	72.36	98.20	100.00	100.00	62.16	89.87	96.66	20.49	47.97	66.12	97.44	99.95	100.00
20m x 20m	85.85	98.46	99.55	23.27	46.99	58.39	98.89	99.97	100.00	60.01	87.96	92.97	14.69	35.64	48.46	99.42	100.00	100.00
40m x 40m	76.44	96.34	98.89	23.54	50.57	62.18	99.10	100.00	100.00	57.72	86.77	91.16	14.15	34.59	45.00	98.98	100.00	100.00
60m x 60m	68.30	93.32	96.24	14.10	30.69	38.06	98.04	100.00	100.00	47.40	79.28	85.96	8.31	19.48	25.75	98.59	100.00	100.00
80m x 80m	64.22	89.40	93.53	12.06	27.46	34.46	99.73	100.00	100.00	43.98	73.57	81.49	6.58	15.62	21.44	99.87	100.00	100.00
100m x 100m	59.32	87.44	91.60	11.03	23.96	31.20	95.73	100.00	100.00	41.36	71.39	79.09	5.83	14.93	20.98	91.01	100.00	100.00



Query Image Prediction Query Image Prediction

Figure 2. Longitudinal pose estimation ambiguity by ground-to-satellite image matching. Determining the camera's position along the road (longitudinal pose) is hard because the scenes along the driving direction are rather monotonous, but the lateral pose can be easily estimated.

.

6. Weakness Discussions

As discussed above, as the rotation and translation ambiguity increases, the localization performance decreases because the scene content captured by the query image may be similar to the scene contents on different parts of the satellite image. Additionally, there is an inherent ambiguity in longitudinal pose estimation when using a single query image for ground-to-satellite image matching, as shown in Fig. 2. Leveraging a multi-camera system with a 360° field of view or a continuous video could potentially improve the informativeness of the query place and thus improve the localization performance. We will investigate these possibilities in the future.