Supplementary Material for CVLNet: Cross-View Feature Correspondence Learning for Video-based Camera Localization

Yujiao Shi[1], Xin Yu[2], Shan Wang[1], Hongdong Li[1]

[1]Australian National University [2]University of Technology Sydney

1 Dataset Statistics

In this section, we provide more illustrations of the introduced KITTI-CVL dataset. Fig. 1 presents an overview of the sampling distributions of the training and testing sets, where training images are captured from the red area, and testing images are sampled in the blue region. The training and testing sets do not overlap. The Validation set is sampled from the same area of the training set.

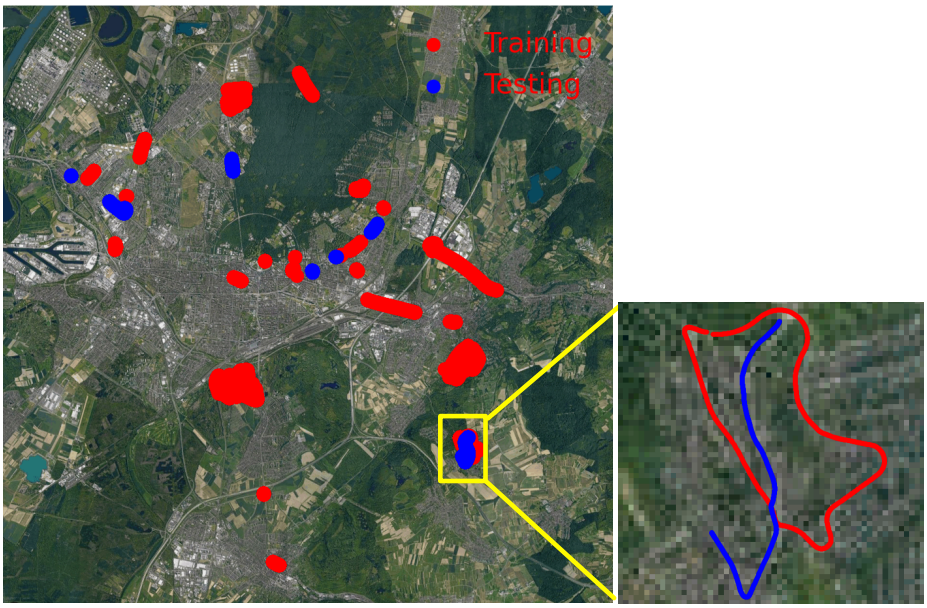
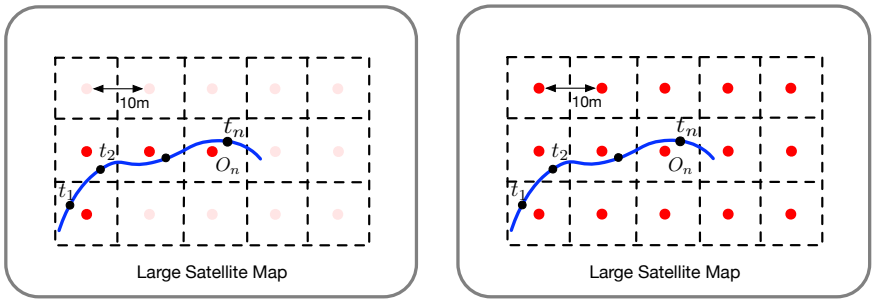


Fig. 1: Training and testing data distribution of the introduced KITTI-CVL. They are captured from different regions.

Differences between Test-1 and Test-2. Our two test sets, *i.e.*, Test-1 and Test-2, share the same query ground images. Their differences lie in satellite



(a) Test-1 (b) Test-2

Fig. 2: Comparison between (a) Test-1 and (b) Test-2. Here, the large black box indicates the whole region of interest. The red dots represent the satellite image centers in the database. They are sampled every ten meters in the region of interest. The blue curve represents a trajectory of a camera. The black dots on the curve denote the camera locations at different time steps. In Test-1, only the nearest satellite image for each query ground image is retained in the database. Thus, we mask the other satellite image centers in the grid, which can be regarded as a non-distractor case. In Test-2, all the satellite images sampled in the grid are reserved.

lat = 49.025527549337 lat = 49.025527483717 lat = 49.025527419798
lon = 8.4485623653485 lon = 8.4485623463895 lon = 8.4485623793197

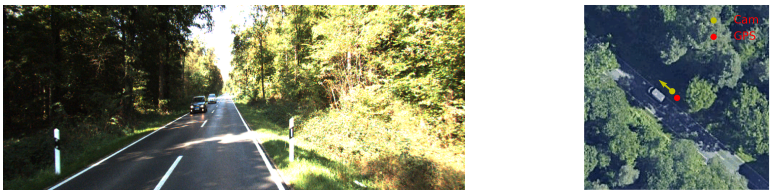
Fig. 3: Varying GPS tags for a static camera. The distances between the tags are from 0.01m to 0.3m. Images are from drive "2011_09_26_drive_0017_sync".

Table 1: Additional ablation study results of our method

Method	Test-1				Test-2			
	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
Ours (GVP on Img)	1.33	5.82	10.23	54.14	0.16	0.36	0.93	13.18
Ours w/o High Objects	2.35	6.67	10.76	52.45	0.65	2.71	3.80	23.49
Ours	**21.80**	**47.92**	**64.94**	**99.07**	**12.90**	**27.34**	**38.62**	**85.00**

images in the database. As shown in the left of Fig. 2, in Test-1, only the nearest satellite image of each query image is retained in the database. While in Test-2, all the satellite images within the large area are reserved. Test-2 has more distracting images in the database and thus is a more challenging test set than Test-1.

GPS noise. We found there is slight noise in the GPS raw data provided by KITTI. For example, as shown in Fig. 4, the ground camera is not moving, while the provided GPS data varies for these images. Fig. 4 presents another example where the camera is expected to be on the road while the camera location provided by the GPS data is on top of the vegetation area. We guess those minor errors make the GPS loss proposed in Zhu *et al.* [51] not work well in the KITTI-CVL dataset.



Ground Image Satellite Image

Fig. 4: The query ground image is captured on the road, while its position from GPS data shows it is on top of a vegetation area. The error is around 1m. In the satellite image, the red point indicates the position of the GPS device on the car of the KITTI dataset, and the yellow point indicates the position of the left color camera. The yellow arrow tags the vehicle heading direction. Images are from drive "2011_09_26_drive_0027_sync".

Ground Image Projected Image Satellite Image

Fig. 5: Visualization of projected images in the overhead-view.

2 Additional Illustrations and Experiments

To better illustrate that GVP is able to project ground-view images following the geometric constraints, we apply the GVP to the original ground-view images, as shown in Fig. 5. It can be seen that pixels on the ground plane have been successfully restored in the overhead view. While for pixels above the ground plane, they undergo obvious distortions. Hence, we apply the GVP to feature level rather than image level in our framework. The high-level features are expected to establish semantic correspondences between the two views and circumvent the side effects of distortions in projection. We provide the performance of applying GVP to image level in the first row of Tab. 1, denoted as "Ours (GVP on Img)". Not surprisingly, it is significantly inferior to Ours, where the GVP is applied to the feature level.

Next, we use semantic maps to filter out the scene objects above the ground plane, *e.g.*, buildings, trees, sky, etc. from the input ground-view images, and feed the processed images to our network, denoted as "Ours w/o High Objects". Fig. 6 shows the comparison between the original ground-view images and the processed images. The performance of "Ours w/o High Objects" is presented

Fig. 6: Comparison between original ground images (left) and semantic-filtered ground images (right).

Table 2: Adding baseline results (sequence=1) for alternative sequence fusion algorithms in Sec. 5.1.2.

		Seq	Test-1				Test-2			
Direct Fusion	Conv2D	1	1.98	5.30	9.50	60.01	7.60	18.48	26.49	72.71
		4	1.25	**5.90**	**10.80**	**65.91**	**8.90**	18.44	**26.61**	**76.51**
	LSTM	1	7.44	22.16	35.06	93.97	3.36	9.87	16.17	64.46
		4	**12.53**	**32.11**	**50.42**	**96.93**	**5.78**	**15.89**	**23.01**	**70.60**
Attention-based Fusion	Conv2D	1	15.37	41.16	56.98	95.79	9.18	20.58	30.53	76.83
		4	**18.80**	**47.03**	**61.75**	**96.64**	**11.69**	**25.03**	**36.55**	**81.52**
	LSTM	1	11.69	37.28	56.65	96.44	7.04	18.16	26.57	79.09
		4	**15.93**	**47.88**	**66.03**	**97.61**	**9.70**	**24.30**	**35.26**	**85.08**
	Ours	1	17.71	44.56	62.15	98.38	9.38	24.06	34.45	78.37
		4	**21.80**	**47.92**	**64.94**	**99.07**	**12.90**	**27.34**	**38.62**	85.00

Table 3: Comparison results of our method with or without satellite image height estimation (sequence = 4).

	Test-1				Test-2			
Ours w/ height estimation	21.57	47.41	65.16	98.30	12.15	26.70	38.72	84.95
Ours w/o height estimation	21.80	47.92	64.94	99.07	12.90	27.34	38.62	85.00

in the second row of Tab. 1. It can be seen that the performance is also worse than ours, indicating that our GVP module on feature level successfully encodes information of scene objects that have higher heights.

We also add the baseline results with a sequence number as one for all the sequence fusion alternatives illustrated in Sec. 5.1.2 in the main paper, as shown in Tab. 2. Note that "Conv3D" is the same as "Conv2D" when sequence=1. The results indicate that using a longer sequence generally helps achieve better performance. Furthermore, we tried to borrow ideas from MVS to estimate overhead-view satellite image height maps from sequential ground view images. The comparison results between with and without height estimation are shown in Tab. 3. There are no significant differences in the final results, but height

Table 4: Comparison results on CVUSA and CVACT (recall at top-1).

CVUSA					CVACT				
SAFA	Polar-SAFA	DSM	Zhu *et al.*	Ours	SAFA	Polar-SAFA	DSM	Zhu *et al.*	Ours
68.03	72.15	63.17	53.77	**72.75**	56.69	62.71	55.07	49.45	**66.30**

Table 5: Comparison with SOTA single image-based localization on video (sequence=4) with average ensemble method.

	Test-1				Test-2			
	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
CVM-NET	2.51	6.71	8.53	9.91	0.04	0.57	1.05	1.58
CVFT	0.08	0.28	1.54	2.67	0.00	0.00	0.00	0.08
SAFA	2.51	5.58	6.15	7.44	0.28	0.53	0.61	0.77
Polar-SAFA	3.84	5.62	6.55	7.36	0.16	0.24	0.44	0.57
DSM	11.04	29.28	42.01	94.90	1.82	6.91	11.12	39.83
Zhu *et al.*	1.78	4.37	5.34	5.86	0.08	0.24	0.44	0.57
Toker *et al.*	4.00	12.70	20.50	79.58	2.39	5.50	8.90	27.05
Ours	**21.80**	**47.92**	**64.94**	**99.07**	**12.90**	**27.34**	**38.62**	**85.00**

Table 6: Comparison with SOTA single image-based localization on video (sequence=4) with majority voting ensemble method.

	Test-1				Test-2			
	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
CVM-NET	6.47	19.33	29.68	65.83	1.09	3.52	6.11	20.30
CVFT	1.42	5.74	12.09	51.19	0.08	0.53	1.74	7.80
SAFA	5.10	14.48	20.50	62.23	1.70	3.64	5.18	15.65
Polar-SAFA	7.48	16.90	25.84	59.97	1.66	2.91	4.73	16.42
DSM	14.27	33.68	45.21	72.26	6.03	13.83	18.56	33.89
Zhu *et al.*	5.34	16.13	23.78	61.26	0.49	1.82	3.40	13.10
Toker *et al.*	4.00	12.70	20.50	79.58	2.39	5.50	8.90	27.05
Ours	**21.80**	**47.92**	**64.94**	**99.07**	**12.90**	**27.34**	**38.62**	**85.00**

estimation introduces more computation and memory. Hence, we do not include height estimation in our final method but use the ground-plane assumption in the projection.

CVUSA & CVACT. Our method applies to panoramas as long as camera parameters are given. However, they are not available in the existing panorama datasets, e.g., CVUSA and CVACT. Furthermore, the matching satellite image centers align precisely with query locations in the two datasets, which is not in practice. To make the experiments meaningful, we (i) approximate the camera parameters of the two datasets by visual and geometry verifications; and (ii) randomly translate satellite images (0-36 pixels) to make their centers not aligned with query locations. Results are shown in Tab. 4. It can be seen our method outperforms SOTA methods. The SOTA results are inferior to that in their original paper because of the practical setting (as in (ii)).

Compare with single image-based localization algorithms on video using ensemble techniques. We adopt two ensemble techniques for single image-based localization on video. The first one is average, and the other one is majority voting. For the majority voting, we compare the distance between every two retrieved results for the images in the video and find the minimum one. The

average position of the two retrieved results whose distance is the minimum is regarded as the query camera location. The results are presented in Tab. 1 and Tab. 2, respectively. No matter which ensemble mechanism is applied, the SOTA single image-based localization methods still achieve inferior results than ours.