View From Above: Orthogonal-View aware Cross-view Localization

Supplementary Material

A. Evaluation of Other FordAV-CVL Dataset Logs

We employed the ‘Log4’ trajectory for method evaluation, following the approach outlined in SIBCL [7]. Additional evaluations on other logs from the FordAV-CVL Dataset are presented in Tab. 1 to complement the results in Tab. 2 of the main paper. For all evaluations, we used ‘2017-08-04-26’ as the training dataset, ‘2017-07-24’ as the validation dataset and ‘2017-10-26’ as the test dataset. Our method consistently estimates accurate 3-DoF poses with low spatial and orientation errors across various scenarios, including freeways ‘Log1’, residential areas ‘Log3’ and ‘Log5’, university campuses ‘Log4’, and vegetated regions ‘Log4’ and ‘Log5.

Table 1. Performance of our method on additional logs of the FordAV-CVL dataset with initial range (±45°,±20m)

		Lateral(m)↓		Lateral(%)↑			Longitudinal(m)↓		Longitudinal(%)↑			Orientation(°)↓		Orientation(%)↑		
	Log	Mean	Median	r@0.25m	r@0.5m	r@1m	Mean	Median	r@0.25m	r@0.5m	r@1m	Mean	Median	r@1°	r@2°	r@4°
1C	Log1	0.37	0.22	55.07	86.58	98.87	0.31	0.22	58.07	98.12	99.41	10.02	3.95	14.07	27.99	50.33
	Log3	0.33	0.28	45.29	77.48	98.97	0.47	0.45	4.47	64.70	99.97	4.42	2.59	18.40	39.71	69.24
	Log5	0.40	0.28	45.49	78.31	97.96	0.43	0.38	19.67	79.09	99.84	7.54	6.27	11.67	22.57	34.89
	All	0.38	0.27	47.35	76.97	97.56	0.35	0.28	40.25	92.14	99.71	6.97	3.43	16.00	32.06	55.61
4C	Log1	0.11	0.10	93.54	99.88	100.00	0.13	0.11	88.29	99.42	100.00	7.76	2.54	25.23	43.62	62.39
	Log3	0.09	0.07	94.83	99.21	100.00	0.17	0.16	77.25	99.44	100.00	3.61	2.89	15.42	32.06	66.00
	Log5	0.09	0.07	98.55	100.00	100.00	0.09	0.08	96.67	100.00	100.00	7.64	6.37	8.24	17.42	32.07
	All	0.10	0.08	94.83	99.37	100.00	0.12	0.10	88.59	99.67	100.00	5.28	3.01	18.98	36.04	60.70

‘Log2’ and ‘Log6’ are excluded due to the construction of roads and buildings during data collection.
‘ALL’: All images from ‘Log1’, ‘Log3’, ‘Log4’, and ‘Log5’.

B. Accumulated Pose Estimation: Qualitative Results on FordAV-CVL Dataset

In this section, we visually present additional trajectories of accumulated pose estimation using the FordAV-CVL dataset. Fig. 1 demonstrates the estimated trajectory using four cameras and Fig. 2 shows that using a front camera.

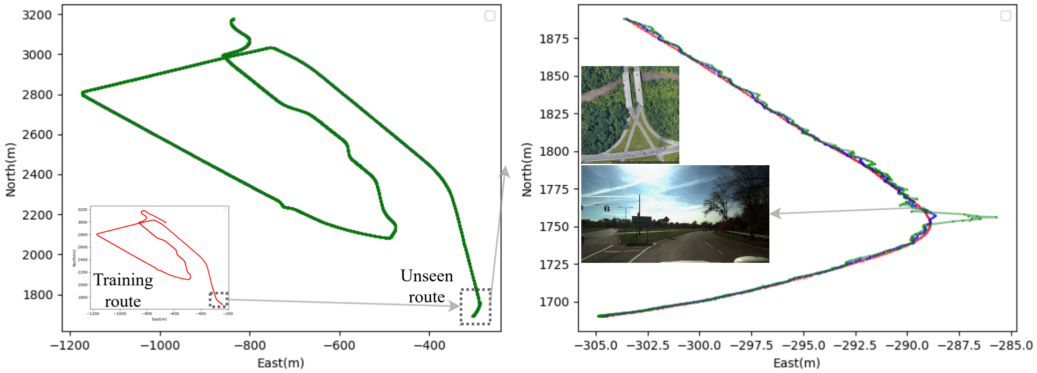

Figure 1. Performance of accumulated pose estimation using four surrounding cameras on the ‘Log4’ trajectory (4189 meters) from the FordAV-CVL dataset. (Left) The predicted poses by our method are illustrated in blue, PureACL’s [6] in green, and the ground truth in red. Both methods successfully estimate the pose throughout the entire route, including the unseen trajectory. (Right) Our method consistently maintains close alignment with the ground truth, even in the unseen parts of the trajectory. In contrast, PureACL experiences some drift but ultimately manages to recover and realign with the correct trajectory.

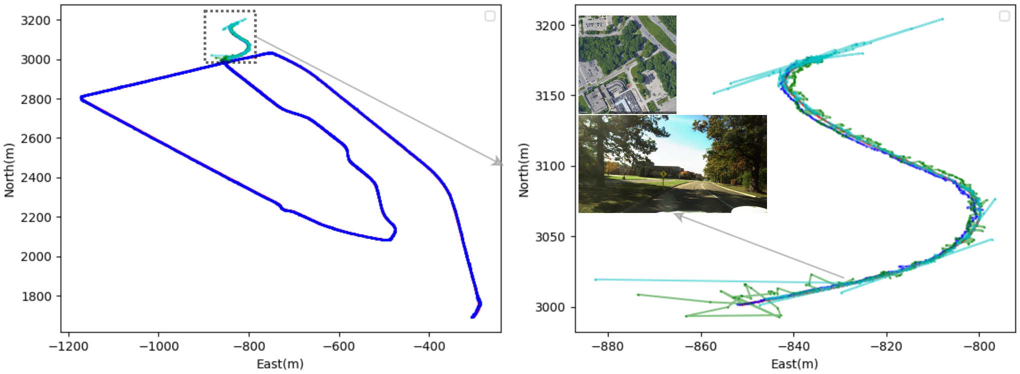

Figure 2. Accumulated pose estimation performance from a front camera on Ford-CVL dataset's 'Log4' trajectory. (Left) Predicted poses by our method (blue) closely track the ground truth (red) along the entire route, outperforming PureACL [6] (green) and BoostAccurate [5] (cyan), which exhibit severe drift. (Right) PureACL and BoostAccurate exhibit significant drift from a similar starting point with severe occlusions, leading to substantial position errors and eventual departure from the satellite map's coverage.

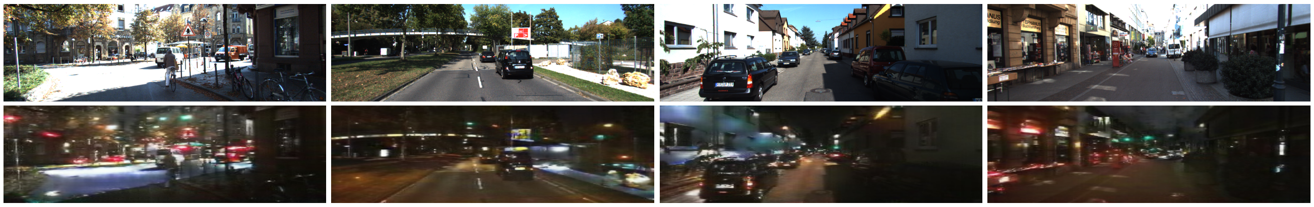

Figure 3. Sample images from the Day-to-Night Dataset. Top: original daytime images from the KITTI dataset. Bottom: synthesized night-time images.

C. Evaluation on an Artificial Day-to-Night Dataset

To thoroughly assess our method's resilience to temporal and lighting changes, we evaluate it on an artificial dataset. This dataset is generated from the KITTI test sets using a cross-domain deep network proposed by Arruda et al. [1], which transforms daytime scenes into their nighttime analogs. This transformation adjusts the images' lighting conditions and introduces nocturnal elements such as artificial lights and reduced visibility. Fig. 3 presents sample images, and Tab. 2 shows that our method maintains consistent performance in nighttime conditions, emphasizing its robustness to changes in light and periods of the day. Such robustness is essential for real-world applications where lighting conditions can vary dramatically, proving the practicality and reliability of our approach in diverse operational environments.

Table 2. Performance of our method on KITTI Day-to-Night Dataset with initial noise (±45°,±20m)

	Model	Lateral(m)↓		Lateral(%)↑			Longit(m)↓		Longit(%)↑			Orient(°)↓		Orient(%)↑		
		Mean	Median	r@0.25m	r@0.5m	r@1m	Mean	Median	r@0.25m	r@0.5m	r@1m	Mean	Median	r@1°	r@2°	r@4°
Same	BoostAcc[5]	2.24	1.13	7.79	17.20	46.77	10.02	8.64	1.41	2.79	5.93	11.22	5.11	10.35	21.64	40.94
	PureACL[6]	5.74	0.43	32.93	57.83	69.62	7.50	1.54	11.34	21.11	37.76	10.46	4.84	12.29	24.43	44.07
	Ours	**0.32**	**0.26**	**47.15**	**80.85**	**98.57**	**0.34**	**0.30**	**37.24**	**86.21**	**99.42**	**4.51**	**1.87**	**29.94**	**52.61**	**75.50**
Cross	BoostAcc[5]	6.08	1.69	9.56	18.16	34.01	10.05	9.24	1.43	2.70	5.72	12.04	8.37	7.5	14.73	26.31
	PureACL [6]	7.38	0.46	28.77	52.65	63.64	8.38	1.77	7.98	16.18	30.97	12.41	5.35	10.21	21.03	39.28
	Ours	**0.39**	**0.27**	**46.33**	**77.78**	**96.96**	**0.39**	**0.30**	**39.33**	**79.26**	**98.30**	**6.64**	**2.55**	**22.86**	**42.29**	**64.18**

D. Evaluation of Robustness to Temporal Illumination Changes

We evaluate our method's resilience to temporal illumination changes by conducting experiments on selected frames that exhibit sudden illumination shifts, following the experimental setup detailed in Sec. 5.4 of main paper. Fig. 4 illustrates our method's robustness to these abrupt changes in illumination.

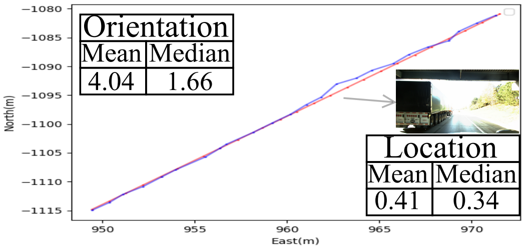

Figure 4. Experiment on illumination changes with 1-camera using frames$_{*87371406.png\sim*89065099.png}$ from "2017-10-26-V2-Log1" in the Ford Multi-AV dataset .

E. Reduced Process Frame Analysis

Although our method does not process every frame in real-time, operating at an approximate speed of 4.5 FPS, it's important to note that by processing once every four frames (resulting in around 3.75 FPS), our approach still preserves its accuracy. As a post-processing pose refiner, our method effectively refines poses and ensures route completion, as demonstrated in Fig. 5. The evaluation, conducted according to the protocol outlined in Sec. 5.4 of main paper, highlights our method's reliability and accuracy even with reduced frame processing.

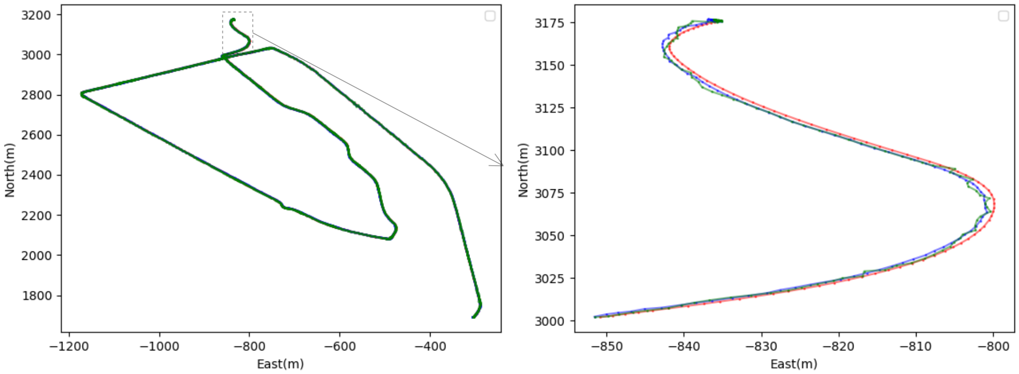

Figure 5. Processing one frame every four frames (3.75 FPS) (green) yields results closely aligned with full-frame (blue) and ground truth (red) throughout the route.

F. Baseline versus Proposed Method

For completeness, we briefly outline the baseline PureACL method [6], as depicted in Fig. 6 (excluding the red text). This method employs a shared-weight Feature Extractor to extract feature maps $F = \{F^s, F^g\}$ from spatially embedded satellite and ground view images $I \copyright E$, where $I = \{I^s, I^g\}$. The embedding E includes three channels containing pixel-wise information on orientation, distance, and height, calculated from the initial pose $\mathbf{P}_{init}$. A Confidence Map Generator produces two types of confidence maps: a view consistent confidence map $V = \{V^s, V^g\}$ and an on-ground confidence map $O = \{O^s, O^g\}$. The former assigns high confidence to objects consistent across both views, while the latter focuses on on-ground objects. In Fig. 6 and our main paper, the overall confidence is represented as $C = V \times O$, integrating both view consistency and on-ground object emphasis. The Keypoint Detector then selects the top-N points p^g with the highest confidence from the ground view confidence map C^g. These points are converted into 3D coordinates p within the vehicle world coordinate system, leveraging their on-ground characteristics. Following this, the Weighted Residual Calculator projects these 3D points onto the satellite view $p^s = K_s \mathbf{P} p$ retrieves the point weight $\{C^s[p^s(\mathbf{P})], C^g[p^g]\}$ and point feature vector $\{F^s[p^s(\mathbf{P})], F^g[p^g]\}$ from the confidence and feature maps, respectively. It then calculates the weighted residual $w(\mathbf{P}) \times r(\mathbf{P})$ as per Eq. 1 in the main paper. Finally, the vehicle's pose $\mathbf{P}_{pred}$ is iteratively refined using the Levenberg-Marquardt (LM) algorithm, which aims to minimize the weighted point residuals.

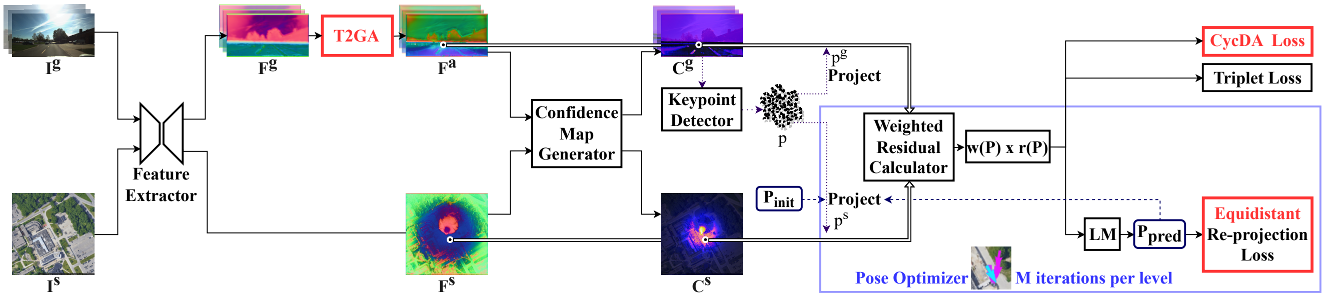

Figure 6. The overall network architecture of our method employs the dual-branch structure of [6], with one branch dedicated to ground views and the other to satellite views. A shared-weight Feature Extractor extracts feature maps $F^{g/s}$ from each input image $I^{g/s}$. The **T2GA** module (Sec. 4.1), highlighted in red, enhances features of on-ground pixels with information from elevated pixels above them. The Confidence Map Generator then produces confidence maps $C^{g/s}$ for each feature map, with ground view confidence maps C^g utilized by the Keypoint Detector to identify high-confidence keypoints. These keypoints are projected into 3D space and subsequently re-projected onto the satellite view by the Weighted Residual Calculator. The Pose Optimizer refines the vehicle's pose $\mathbf{P}_{pred}$ using the Levenberg-Marquardt algorithm over M iterations at each level, from coarse to fine. Our architecture integrates the **CycDA Loss** (Sec. 4.2) and Triplet Loss for view-invariant feature representation, along with the **Equidistant Re-projection Loss** (Sec. 4.3), which normalizes the influence of distant keypoints and encourages the predicted pose $\mathbf{P}_{pred}$ to closely align with the ground truth $\mathbf{P}_{gt}$.

The pose optimization is performed iteratively using the equation below:

$$\delta_{t+1} = \delta_t - (\mathbf{H} + \lambda\, \text{diag}(\mathbf{H}))^{-1}\mathbf{J}^\top \mathbf{W}\Upsilon, \tag{1}$$

where δ represents an individual element in the 3-DoF pose. The current iteration is represented by $t \in \{1, \cdots, M \times L\}$, where M is the iteration count per level and L is the total number of levels. The damping factor λ follows the definition in [4]. The matrices $\mathbf{W} \in \mathbb{R}^{NN}$ and $\Upsilon \in \mathbb{R}^{ND}$ are constructed by stacking all point weights $\mathbf{W} = \text{diag}(w(\mathbf{P})\rho')$ and residuals $r(\mathbf{P})$, respectively, with ρ' being the derivative of the robust cost function [2]. The Jacobian $\mathbf{J}$ and Hessian matrices $\mathbf{H}$ are defined as shown below:

$$\mathbf{J} = \frac{\partial r(\mathbf{P})}{\partial \delta} = \frac{\partial F^s[p^s(\mathbf{P})]}{\partial p^s(\mathbf{P})}\frac{\partial p^s(\mathbf{P})}{\partial \delta}, \quad \mathbf{H} = \mathbf{J}^\top \mathbf{W}\mathbf{J}. \tag{2}$$

PureACL [6] employs two loss functions: **Re-projection Loss**, which encourages the predicted pose $\mathbf{P}_{pred}$ to closely align with the ground truth $\mathbf{P}_{gt}$, and **Triplet loss** [3], designed to enhance pose-sensitive feature extraction. This is achieved by reducing the overall error of weighted residuals at the ground truth pose $\varepsilon(\mathbf{P}_{gt})$ and penalizing this error at incorrect initial poses $\varepsilon(\mathbf{P}_{init})$. The form of the Re-projection Loss is detailed as per Eq. 8 in the main paper, while the formulation of the Triplet Loss is as follows::

$$L_{TRP} = log(1 + e^{\alpha(1-\frac{\varepsilon(\mathbf{P}_{init})}{\varepsilon(\mathbf{P}_{gt})})}), \quad \varepsilon(\mathbf{P}) = \textstyle\sum_p w(\mathbf{P}) \times \rho(\|r(\mathbf{P})\|_2^2), \tag{3}$$

where α is a margin-controlling hyper-parameter, the summation $\sum_p$ aggregates over all keypoints, and $\rho(\cdot)$ denotes a robust cost function [2].

Our method introduces three novel components, highlighted in red in Fig. 6: (1) **T2GA** (Sec. 4.1), (2) **CycDA Loss** (Sec. 4.2), and (3) **Equidistant Re-Projection Loss** (Sec. 4.3), each grounded in key insights:

- Vertical landmarks, such as trees, streetlights and traffic lights are reliable features across ground and satellite images. Making use of such landmarks for cross-view matching will enhance localization accuracy. To this end, **T2GA** is developed to make use of these vertical landmarks by augmenting on-ground pixels with the elevated pixels directly above them, thereby reducing the representation gap of the same object across different views.
- Ground and satellite images exhibit significant appearance differences due to varying view angles, sensors, and environmental conditions like weather, lighting, and timing. These differences can introduce significant bias and need to be treated properly for robust cross-view matching. To address this, we propose **CycDA Loss**, which enforces a view-invariant representation for identical objects across different domains.

- Finally, ground and satellite images differ fundamentally in their image formation methods: perspective projection for ground images and quasi-parallel projection for satellite images. In perspective projection, areas near the horizon line are compressed compared to their representation in satellite images. This compression results in those areas contributing less to the optimization process, despite their larger representation and potential importance for orientation sensitivity in satellite images. The proposed **Equidistant Re-Projection Loss** addresses this by distributing weights more uniformly, enabling a more extensive ground region to be matched across views and thereby enhancing orientation accuracy.

These components, informed by these insights, contribute significantly to advancing more robust and accurate cross-view localization solutions.

References

[1] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2

[2] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011. 4

[3] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6450–6458, 2019. 4

[4] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021. 4

[5] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer, 2023. 2

[6] Shan Wang, Yanhao Zhang, Akhil Perincherry, Ankit Vora, and Hongdong Li. View consistent purification for accurate cross-view localization, 2023. 1, 2, 3, 4

[7] Shan Wang, Yanhao Zhang, Ankit Vora, Akhil Perincherry, and Hengdong Li. Satellite image based cross-view localization for autonomous vehicle. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3592–3599. IEEE, 2023. 1