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

Shan Wang[1,2] Chuong Nguyen[1] Jiawei Liu[2] Yanhao Zhang[2]
Sundaram Muthu[1] Fahira Afzal Maken[1] Kaihao Zhang[2] Hongdong Li[2]
[1]Data61, CSIRO [2]Australian National University

Abstract

This paper presents a novel aerial-to-ground feature aggregation strategy, tailored for the task of cross-view image-based geo-localization. Conventional vision-based methods heavily rely on matching ground-view image features with a pre-recorded image database, often through establishing planar homography correspondences via a planar ground assumption. As such, they tend to ignore features that are off-ground and not suited for handling visual occlusions, leading to unreliable localization in challenging scenarios. We propose a Top-to-Ground Aggregation (T2GA) module that capitalizes aerial orthographic views to aggregate features down to the ground level, leveraging reliable off-ground information to improve feature alignment. Furthermore, we introduce a Cycle Domain Adaptation (CycDA) loss that ensures feature extraction robustness across domain changes. Additionally, an Equidistant Re-projection (ERP) loss is introduced to equalize the impact of all keypoints on orientation error, leading to a more extended distribution of keypoints which benefits orientation estimation. On both KITTI and Ford Multi-AV datasets, our method consistently achieves the lowest mean longitudinal and lateral translations across different settings and obtains the smallest orientation error when the initial pose is less accurate, a more challenging setting. Further, it can complete an entire route through continual vehicle pose estimation with initial vehicle pose given only at the starting point. [1]

1. Introduction

Visual-based cross-view localization aims to locate query images taken from street-level cameras (referred to as ground view) within a satellite or aerial view map. Platforms like the Google Map API [6] have made satellite images accessible, spurring the development of cross-view localization techniques. Notable studies [4, 7, 10, 15, 20, 25, 26, 30] have focused on using satellite images for this task. Despite this, accurate localization remains challenging due to the significant viewpoint differences between ground and aerial images [14, 31]. These viewpoint differences result in a domain gap, adversely impacting feature alignment, thereby compromising the overall accuracy of localization.

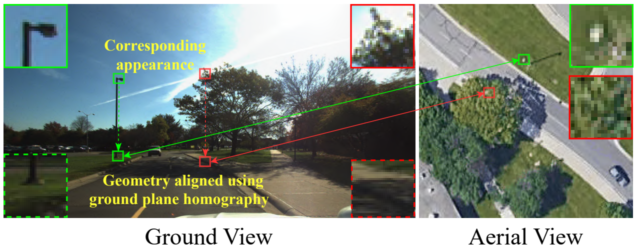


Figure 1. Efficacy of T2GA. T2GA aggregates **off-ground features** (e.g., the streetlight within the green square) and addresses **occlusions** (e.g., the tree within the red square). Unlike conventional methods that often overlook such elements, T2GA integrates these features, thus improving appearance matching for ground-level aligned pixels across views using ground plane homography. Insets are magnified for clarity, with dotted outlines indicating appearance misalignment and solid outlines indicating appearance alignment with the corresponding aerial view features.

Recent research has explored two main approaches to bridge the domain gap in image-based localization: generative-based and geometry-alignment-based methods. Generative-based methods, such as those utilizing GANs [19] and diffusion models [29], reduce the domain gap by transforming view styles from one view to another. However, the generated features for matching can lead to ambiguities in pose estimation. In contrast, geometry-alignment-based methods, employing techniques like polar transformations [15, 20] or homography [14, 23, 24, 26], focus on establishing correspondences for on-ground pixels. This often results in the neglect of **off-ground features**, such as streetlights, and difficulty in handling visual **occlusions**, such as the obscuring of road details by treetops in aerial views, both illustrated in Fig. 1. This neglect fails to leverage important geographic landmarks on the road and results in a lack of robustness to issues like road mark degradation (*e.g.* fading and damaged paintings). To address these limitations, we introduce the Top-to-Ground Aggregation (T2GA) module. T2GA employs top-down feature aggre-

[1] Code is available at https://github.com/ShanWang-Shan/ViewFromAbove.

gation to enrich on-ground points with an above-view perspective appearance, significantly improving feature alignment and localization accuracy. Furthermore, we recognize that pixel positions above ground points may not necessarily indicate higher elevations; instead, they could represent objects located further along the camera's line of sight. To resolve this ambiguity, we integrate a transformer mechanism that assesses whether these pixels belong to the same object and determines occlusion precedence, such as shadows on the road, which suggest likely occlusions in the aerial view.

In addition to viewpoint differences, ground and aerial views differ in cameras types, lighting conditions, tones, and resolutions. We introduce the Cycle Domain Adaptation (CycDA) loss function to address these variations. CycDA enables bidirectional feature generation between ground and aerial views. By minimizing the discrepancy between domain-adapted features and their target counterparts, our approach ensures that features extracted from one domain are effectively translatable to the other, fostering the extraction of features that are invariant across domains.

We further introduce an Equidistant Re-projection Loss to address a common bias in keypoint-synchronous detection localization methods, which tend to favor closer keypoints due to their reduced orientation errors. Our loss function mitigates this issue by applying a distance-weighted approach, ensuring that orientation errors are independent of keypoint distance. Consequently, keypoints are more uniformly distributed across various distances, leading to a more equitable and precise estimation of direction.

We summarize our contributions as follows:

- A top-to-ground feature aggregation module that effectively bridges the domain gap between ground and aerial views by utilizing off-road cues and handling occlusions.
- A cycle domain adaptation loss function that promotes domain-invariant feature extraction, enhancing the robustness of cross-view localization methods.
- An equidistant re-projection loss function ensures orientation errors are consistent regardless of keypoint distance, leading to a more extended distribution of keypoints and more accurate orientation estimation.

The evaluations are conducted on KITTI and Ford Multi-AV datasets while leveraging the Google Maps as corresponding satellite images. Experimental results demonstrate that our method can achieve consistent vehicle pose estimation under various challenging situations. Our method achieves the lowest mean longitudinal and lateral translation errors on both KITTI and Ford Multi-AV datasets under different settings. We also obtains the smallest orientation error when the initial pose is less accurate, demonstrating the robustness of our method. Further, with initial vehicle pose only at the starting point, our method can complete the route through continuous pose estimation, demonstrating its generalisation ability and potential for practical deployment.

2. Related Work

Image-level cross-view localization. Early visual-only cross-view localization methods [7, 10, 15, 16, 20, 30] approach the task as an image retrieval problem, focusing on coarse localization through image-to-image matching. To bridge the domain gap between ground and aerial views, various techniques have been proposed to facilitate cross-view feature matching. [10] incorporated per-pixel orientation information, while [15] and [20] utilized a predefined polar transform to align aerial-view images with ground views. Additionally, [19] and [29] employed GAN-based and diffusion model-based style transformations, respectively. While efforts have been made to minimize the domain gap, these methods rely on image-level feature matching, restricting their localization accuracy, often falling short of the precision achieved by commercial GPS systems in open areas [21].

Patch-wise cross-view localization. To improve accuracy, several methods employ patch-wise feature matching, especially effective in aerial views with wide fields of view and high resolutions. For instance, [31] employs transformers to emphasize informative patches, [27, 28] computes dense spatial distributions using patch attention, and [9] introduces the 'slice-to-sector match' and 'Cross-view attention' to calculate similarity between aerial sectors and ground view slices. However, their attention queries are derived from aerial images, posing reliability issues when the pose is unknown. OrienterNet [13] transforms ground view images into Bird's Eye View (BEV) grids for matching with OpenStreetMap data. Despite these advancements, localization accuracy remains limited by the patch (grid) size.

Pixel-wise cross-view localization. Pixel-wise feature matching has been explored in various methods [4, 14, 24, 25] for precise localization, yet they struggle with the inherent domain gap between views. Our baseline, PureACL[24], and other geometry-alignment-based methods [14, 26] prioritize on-ground pixel correspondences but overlook off-ground features and occlusions, leading to suboptimal performance. CVGL [4] transforms the ground view image into a BEV for aerial matching. Their BEV transformation employs aerial coordinates as queries, introducing a high degree of freedom and increasing the complexity of matching. In the BoostAcc [17] framework, homography transformation is used on ground views to source query pixel data for the pixel-to-slice attention mechanism. This transformation introduces distortions from non-ground pixels, potentially exacerbating errors in subsequent processing stages. Moreover, the keys and values span entire and adjacent columns, risking compounded ambiguities in both longitudinal and lateral estimation. Our approach aims to address these challenges, proposing a novel pixel-to-overhead-column attention mechanism for improved cross-view localization accuracy.

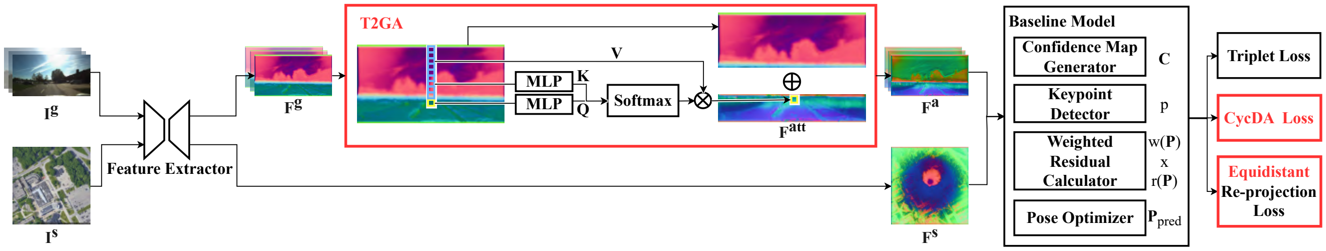


Figure 2. Our method adopts PureACL [24] as the baseline model and introduces three novel components (highlighted in red): (1) **T2GA** (Sec.4.1): aggregates the features of elevated pixels onto the feature of the on-ground pixel that is directly beneath them to alleviate the representation gap of the same object across different views; (2) **CycDA Loss** (Sec.4.2): explicitly enforces a view-invariant representation for the same object; and (3) **ERP Loss** (Sec.4.3): allows the model to leverage key points that are farther away from the vehicle while allocating more emphasis on correcting the vehicle orientation estimation.

3. Preliminary

Task Settings. The Fine-Grained Cross-View Localization (FGCVL) task aims to estimate the accurate 3-DoF pose of a vehicle, including longitudinal and lateral translations and orientation on a satellite map I^s, given an initial coarse pose $\mathbf{P}_{init}$. It also assumes that the vehicle is equipped with on-board camera(s) [2] , and the image captured by these cameras, denoted as the ground image I^g, is accessible.

Baseline Model. PureACL [24] is adopted as the baseline model for our approach. It employs a share-weight U-Net to extract ground/satellite feature maps $F = \{F^g, F^s\}$ and ground/satellite confidence maps $C = \{C^g, C^s\}$ [3] from the ground/satellite images $I = \{I^g, I^s\}$. The ground confidence map C^g is tasked to select the top-N most confident pixel-level features from F^g at the positions $p^g = \{(u^g_i, v^g_i)\}^N_{i=1}$. These features are then: (i) Projected from on-board camera coordinates to the satellite coordinates using $p^s_i(\mathbf{P}) = K_s \mathbf{P} K_g^{-1} p^g_i$ [4], where K_s and K_g are the intrinsic matrices of satellite and on-board cameras, respectively. $\mathbf{P}$ is the initial pose $\mathbf{P}_{init}$ at the first iteration and the estimated pose $\mathbf{P}_{pred}$ in the subsequent iterations. (ii) A weighted residual $w(\mathbf{P}) \times r(\mathbf{P})$ is computed between the top-N selected features from the ground feature map and the corresponding features from the satellite feature map as:

$$w(\mathbf{P}) = C^s[p^s(\mathbf{P})] \times C^g[p^g], r(\mathbf{P}) = F^s[p^s(\mathbf{P})] - F^g[p^g], \quad (1)$$

where $[\cdot]$ is a lookup operation in a feature/confidence map with sub-pixel interpolation. (iii) The Levenberg-Marquardt (LM) algorithm [11] takes the computed weighted residual $w(\mathbf{P}) \times r(\mathbf{P})$ as input and outputs the predicted vehicle pose $\mathbf{P}^m_{pred}$ at the m^{th} iteration, with $m \in \{1, \dots, M\}$. The process of (i) - (iii) repeats for $M = 20$ iterations to produce the final vehicle pose $\mathbf{P}_{pred}$.

PureACL [24] employs two loss functions: a Triplet Loss [12] and a Re-projection Loss. The former supervises the weighted residual, enforcing that the encoder extracts pose-sensitive features from both ground and satellite images. The later utilises the projection error w.r.t the ground truth pose $\mathbf{P}_{gt}$ to penalize incorrectly predicted vehicle pose $\mathbf{P}_{pred}$ (Refer to Supplementary Sec.D for more details).

4. Method

We propose a novel View From Above (VFA) method to tackle the FGCVL task by aligning feature representations across the orthogonal viewpoints, the ground view and the satellite views. As illustrated in Fig. 2, our method introduces three innovative components onto the baseline model: (1) Top-to-Ground Aggregation Module (T2GA) (Sec. 4.1); (2) Cycle Domain Adaptation Loss (CycDA) (Sec. 4.2); and (3) Equidistant Re-Projection Loss (Sec. 4.3).

4.1. Top to Ground Aggregation

The appearance discrepancy between the same objects viewed from ground and satellite perspectives presents a significant challenge in achieving high precision in the FGCVL task [14, 17, 24, 26]. This issue is often encountered on tall structures, *e.g.* street light, and the occlusion from aerial objects, *e.g.* tree branches, where their near(on)-ground appearance, accessible to only on-board vehicle camera, is significantly different from their top appearances which are only registered by the satellite camera (see Fig. 1). Such discrepancies are common on roads and can lead to inaccuracies in cross-view localization. To address this challenge, we propose aligning the representations between ground and satellite views, focusing specifically on the on-ground pixels, where geometry alignment across the two views hold due to the ground plane homography. Specifically, we present a T2GA module that aggregates the features of elevated pixels onto the feature of the on-ground pixel that is directly beneath them, as illustrated in the 'T2GA' red box in Fig. 2. For an on-ground pixel:

[2] While our method is compatible with multiple cameras, our primary experiments focus on one front camera and four surrounding cameras.

[3] We simply refer to the fused confidence map, rather than distinguishing between view-consistent and on-ground confidence maps in PureACL.

[4] PureACL obtains 3D world coordinates of the vehicle from homogeneous coordinates $K_g^{-1} p^g$ by utilizing on-ground characteristics. For simplicity in this context, we note homogeneous coordinates as directly corresponding to the 3D world coordinates.

$p \in \{(u, v)\}_{u=0, v=\tau}^{W^g-1, H^g-1}$, and where τ is a threshold corresponding to the height of horizon in the ground view image, and W^g and H^g are the width and height of the ground view image, respectively. The corresponding elevated pixels are $q \in \{(u_p, v)\}_{v=0}^{v_p}$, where (u_p, v_p) are the coordinates of the on-ground pixel p. The aggregation is achieved via the attention mechanism [22], defined as:

$$F^{att}[p] = \text{Softmax}\ (\mathbf{Q}\mathbf{K}^T)\mathbf{V}, \tag{2}$$

where $F^{att}[p]$ represents the feature vector at a pixel p on the feature map F^{att}. In this context, p and q denote an on-ground pixel and its corresponding elevated pixels, respectively. The query is formed by taking the features of on-ground pixels $\mathbf{Q} = \mathcal{M}(F^g[p])$ where $\mathcal{M}(\cdot)$ represents a non-linear mapping consisting of an MLP layer followed by an activation function. The value is derived from the features of elevated pixels corresponding to the on-ground pixel $\mathbf{V} = F^g[q]$, and the key is then obtained by applying the same non-linear mapping to $\mathbf{V}$, resulting in $\mathbf{K} = \mathcal{M}(\mathbf{V})$. This non-linear mapping ensures that the query and key are mapped to the same feature space before computing their matrix product. $\mathbf{Q}\mathbf{K}^T$ [5]. Notably, value $\mathbf{V}$ is used directly in its original form for the computation, facilitating a straightforward fusion (for the same object) or replacement (in cases of occlusion) at ground level.

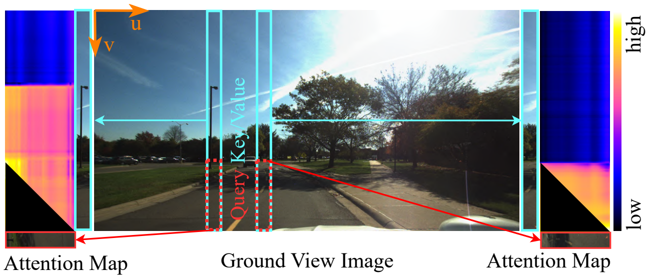


Figure 3. Attention between the on-ground pixels and their corresponding elevated pixels is displayed column-wise. (Left example): The attention between the base and top of the streetlight is high despite their distinct appearances. This allows the resultant aggregated ground feature to be aligned with the satellite feature corresponding to the matching geographic location. (Right example): In the absence of occlusion from above, ground pixels maintain minimal attention with their elevated pixels. This avoids unnecessary dilution of the ground features which are already well aligned with their satellite counterparts.

The computed attention map Softmax$(\mathbf{Q}\mathbf{K}^T)$ can accurately reflect the connection between the on-ground pixels and their corresponding elevated pixels. As illustrated in Fig. 3, high attention values are assigned to the base and top of the streetlight despite their different appearances. This provides important cues to alleviate the representation gap between the ground feature and satellite features. On the other hand, in the absence of occlusion, *e.g.* tall structure and tree branch, thus the attention between on-ground and elevated pixels is minimal, avoiding diluting the ground features that are well aligned with their satellite counterparts.

The feature map F^{att} after attention calculation is vertically stacked with the upper part of the original F^g to generate the aggregated feature map F^a:

$$F^a = F^g_{0:H^g-\tau} \oplus F^{att}, \tag{3}$$

where $\oplus$ denotes vertically stacking and $F^g_{0:H^g-\tau}$ indicated the slicing of F^g from row 0 to $H^g - \tau$. Subsequently, this aggregated feature map F^a replaces F^g in the residual calculation as per Eq. 1, resulting in an updated formula: $r(\mathbf{P}) = F^s[p^s(\mathbf{P})] - F^a[p^g]$.

The advantage of the aggregated feature map is also indirectly reflected by the ground confidence map of the baseline model as shown in Fig. 4. Without the T2GA module, the confidence map only highlights pixels corresponding to road marks and curds, resulting in a concentrated sampling of key points for subsequent estimation of vehicle pose. This has the following drawbacks: (1) fail to leverage road landmarks *e.g.* traffic signal poles, that encode important geographic location; and (2) not robust to road mark degradation problems, *e.g.* fading and damaged paintings, and visibility issues, *e.g.* glow on road marks. T2GA alleviates the feature discrepancy between ground and satellite views. As a result, the traffic signal pole receives higher confidence, allowing it to be subsequently sampled as key points to estimate more precise vehicle pose.

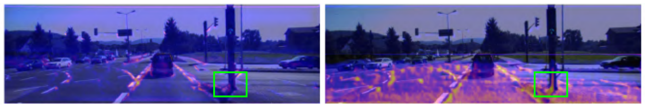


Figure 4. Illustration on the effect of T2GA on the confidence map of the baseline model. The confidence map without T2GA (Left) predominantly highlights road marks and curbs, resulting in subsequent keypoint sampling missing important road landmarks, *e.g.* traffic signal poles, that provide important cues to vehicle pose estimation; The confidence map with T2GA (Right) has high-confidence values distributed across various road marks and traffic poles. With more geographic cues provided by multiple sources, the resultant pose prediction becomes more precise and robust.

4.2. Cycle Domain Adaptation Loss

There are multiple sources responsible for the representation gap between ground and satellite features. In addition to the varying appearances of the same object across different views, camera specifications, *e.g.* tone, hue, intensity, brightness and resolution, and temporal changes of

[5] In processing all query pixels, we first compute the matrix multiplication $\mathbf{Q}\mathbf{K}^T$ using all column pixels $q \in \{(u_p, v)\}_{v=0}^{H^g-1}$. We then eliminate the attention values corresponding to pixels below each query pixel $\{(u_p, v)\}_{v=v_p+1}^{H^g-1}$. After this elimination, the Softmax function is applied.

varying spans, can also lead to inconsistent representations for the same geographic location. Such issue is in general overlooked by the existing FGCVL methods [14, 28]. Despite introducing a triplet loss to further distinguish feature representations based on different poses, our baseline model, PureACL [24], falls short in enforcing invariant feature representation at the corresponding geographic locations across different views.

We propose a Cycle Domain Adaptation (CycDA) loss to explicitly enforce view-invariant representations that consists of three L2-loss between ground and satellite feature representations in three different feature spaces. The representation loss in the ground feature space is defined as:

$$\mathcal{L}_g = \frac{1}{N}\sum_{i=1}^{N}\left\|F^a[p_i^g] - \mathcal{D}\Big(\mathcal{E}\big(F^s[p_i^s(\mathbf{P}_{gt})]\copyright c_s\big)\copyright c_g\Big)\right\|_2, \quad (4)$$

where $\mathcal{E}(\cdot)$ is a projection from the pixel feature space to a latent feature space, $\mathcal{D}(\cdot)$ is a projection from the latent feature space to the pixel feature space, c_s and c_g represent the focal length of satellite camera and vehicle camera respectively which are introduced to guide the projection functions, and $\copyright$ is a concatenation acting on the feature channel dimension. Similarly, the representation loss in the satellite feature space is:

$$\mathcal{L}_s = \frac{1}{N}\sum_{i=1}^{N}\left\|F^s[p_i^s(\mathbf{P}_{gt})] - \mathcal{D}\Big(\mathcal{E}\big(F^a[p_i^g]\copyright c_g\big)\copyright c_s\Big)\right\|_2, \quad (5)$$

We also enforce the satellite feature to be aligned with the ground feature corresponding to the same geographic location in a latent feature space through:

$$\mathcal{L}_m = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{E}(F^a[p_i^g]\copyright c_g) - \mathcal{E}\big(F^s[p_i^s(\mathbf{P}_{gt})]\copyright c_s\big)\right\|_2. \quad (6)$$

The overall CycDA loss can then be defined as:

$$\mathcal{L}_{CycDA} = \mathcal{L}_g + \mathcal{L}_s + \mathcal{L}_m \quad (7)$$

4.3. Equidistant Re-projection Loss

Our method tackles a challenging setting of simultaneously detecting key points from the ground image and predicting the vehicle pose using the detected key points. Our baseline model introduces a re-projection loss computed over the selected top-N key points defined as:

$$L_{RP} = \frac{1}{N}\sum_{i=1}^{N}\|p_i^s(\mathbf{P}_{pred}) - p_i^s(\mathbf{P}_{gt})\|_2^2, \quad (8)$$

where $p_i^s(\mathbf{P}_{pred})$ and $p_i^s(\mathbf{P}_{gt})$ are the projected coordinates on the satellite image from the selected top-N ground key points with the predicted vehicle pose and the groundtruth vehicle pose respectively, and $\|\cdot\|_2$ denotes the L2-distance.

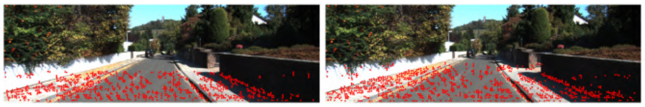


Figure 5. Comparison of detected keypoints with and without ERP Loss. (Left) Without the ERP Loss, keypoints are predominantly located in close proximity to the vehicle. (Right) With the ERP Loss, there is a more dispersed distribution of keypoints.

Despite being effective in pipelines where simultaneous detection of key points is not required [3, 18, 25], the reprojection loss tends to overly penalize the orientation errors of distant points. This forces the detected key points being closer to the vehicle, as illustrated in Fig. 5, and less accurate vehicle orientation estimation (see Sec. 5.2).

To mitigate the aforementioned issue, we propose an Equidistant Re-Projection (ERP) loss that is defined as:

$$L_{ERP} = \frac{1}{N}\sum_{i=1}^{N}\|\frac{p_i^s(\mathbf{P}_{pred}) - p_i^s(\mathbf{P}_{gt})}{\mathbf{D}_{p_i^s}}\|_2^2, \quad (9)$$

where $\mathbf{D}_{p_i^s} = \|p_i^s(\mathbf{P}_{gt}) - p_{vehicle}^s(\mathbf{P}_{gt})\|_2$ is the L2-distance between the satellite image coordinates of the i^{th} keypoint and the vehicle. The design prevents the orientation error from scaling with the L2-distance between key points and vehicle on the satellite image, which dominates over the orientation error on distant key points. As shown in Fig. 5, our method can leverage key points that are farther from the vehicle than our baseline model.

5. Experiments

5.1. Implementation Details

Datasets. To evaluate our proposed approach, we conduct experiments on two well-established autonomous driving datasets: the Ford Multi-AV Seasonal dataset (FMAVS) [1] and the KITTI dataset [5]. Consistent with established practices [14, 24, 25], our primary focus is on images from the front left camera as query images. Additionally, following the methodology of PureACL [24], we broadened our analysis to include four camera perspectives: front left, rear right, side left, and side right. For data partitioning in the KITTI-CVL dataset, we utilized an approach in line with HighlyAcc [14], which involves two test splits. The first, 'Same,' includes images from the same trajectories as the training dataset, while the second, 'Cross,' comprises images from distinct trajectories. Regarding the FMAVS dataset, we follow the partitioning strategy proposed by PureACL [24], using all images from the same trajectories but different traversals. It is noteworthy that there are minor route variations within these same trajectories [6]

Metrics. We adopt evaluation metrics in line with HighlyAcc [14], SliceMatch [9], SIBCL [25], and PureACL [24].

[6] An example is provided in Supplementary Fig. 1.

Table 1. Comparison on the KITTI-CVL Dataset Under Initial Noise Conditions (±10°,±20m). Results marked with ⋆ are sourced directly from the respective original papers. Those denoted by ⋄ indicate retraining of methods for consistent evaluation criteria alignment. - indicates that the corresponding data were not provided.

		Location(m)↓		Lateral(%)↑			Longit(%)↑			Orient(°)↓		Orient(%)↑		
	Model	Mean	Median	r@1m	r@3m	r@5m	r@1m	r@3m	r@5m	Mean	Median	r@1°	r@3°	r@5°
Same	⋆DSM[15]	-	-	10.12	30.67	48.24	4.08	12.01	20.14	-	-	3.58	13.81	24.44
	⋆HighlyAcc[14]	12.08	11.42	35.54	70.77	80.36	5.22	15.88	26.13	3.72	2.83	19.64	51.76	71.72
	⋆SliceMatch[9]	7.96	4.39	49.09	-	98.52	15.19	-	57.35	4.12	3.65	13.41	-	64.17
	⋆BoostAcc[17]	10.01	5.19	76.44	96.34	98.89	23.54	50.57	62.18	0.55	0.42	**99.10**	**100.00**	**100.00**
	⋆CCVPE [28]	1.22	0.62	97.35	98.65	99.71	77.13	96.08	97.16	0.67	0.54	77.39	99.47	99.95
	⋆HC-Net [26]	0.80	0.50	99.01	-	99.73	92.20	-	99.25	**0.45**	**0.33**	91.35	-	99.84
	⋄PureACL [24]	2.42	0.42	91.95	93.40	94.28	91.86	92.12	92.38	3.97	1.77	32.71	53.89	71.66
	Ours	**0.20**	**0.17**	**99.97**	**99.97**	**99.97**	**99.97**	**99.97**	**99.97**	1.53	0.93	50.95	83.97	95.45
Cross	⋆DSM[15]	-	-	10.77	31.37	48.24	3.87	11.73	19.50	-	-	3.53	14.09	23.95
	⋆HighlyAcc[14]	12.58	12.11	27.82	59.79	72.89	5.75	16.36	26.48	3.95	3.03	18.42	49.72	71.00
	⋆SliceMatch[9]	13.50	9.77	32.43	-	86.44	8.30	-	35.57	4.20	6.61	46.82	-	46.82
	⋆BoostAcc[17]	13.01	9.06	57.72	86.77	91.16	14.15	34.59	45.00	**0.56**	**0.43**	**98.98**	**100.00**	**100.00**
	⋆CCVPE [28]	9.16	3.33	44.06	81.72	90.23	23.08	52.85	64.31	1.55	0.84	57.72	92.34	96.19
	⋆HC-Net [26]	8.47	4.57	75.00	-	97.76	58.93	-	76.46	3.22	1.63	33.58	-	83.78
	⋄PureACL [24]	6.20	0.61	67.24	87.26	92.76	64.81	73.63	89.69	4.26	2.48	23.45	48.69	59.24
	Ours	**0.21**	**0.18**	**99.92**	**99.96**	**99.97**	**99.91**	**99.91**	**99.91**	2.04	1.38	38.31	80.12	92.48

Our precision assessment includes median and mean errors for overall, lateral, and longitudinal translations, as well as orientation accuracy. Additionally, our analysis covers lateral and longitudinal translations and localization recall at various distances (0.25m, 0.5m, 1m, 3m, and 5m), and orientation recall within a range of 1° to 5°.

Training Details. We apply two sets of criteria: (1) Following HighlyAcc [14], ground images are processed at a resolution of 256×1024, and satellite images at 512×512, with initial pose noise ±10° for orientation and ±20m for translations. (2) In accordance with PureACL [24], we process FMAVS images at 432×816, KITTI images at 384×1248, and satellite images at $1,280 \times 1,280$. To better simulate turning scenarios, we adopt an expanded initial pose noise range of ±45° for orientation and ±20m for translations. For training, we utilize an NVIDIA RTX 3090 GPU with a batch size of 3, employing the Adam optimizer [8] with a learning rate of 10^{-5}. Feature extractor weights are adapted from PureACL [24], while other components are initialized randomly. Training iterations average around 285ms, including 200ms dedicated to optimization. The inference speed, subject to initial pose variability, averages at 222ms.

5.2. Comparison with Existing Methods

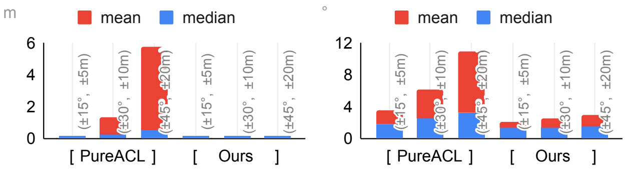
We evaluate our method against recent visual-only approaches on the KITTI-CVL dataset following the metrics of HighlyAcc [14]. The results shown in Tab. 1 clearly indicate the superiority of our approach in spatial precision, consistently maintaining poses within a 1m radius in both 'Same' and 'Cross' areas with over $99.9^{+}\%$ probability. While our method may not lead in orientation accuracy under an initial pose range of ±10°, it demonstrates superior performance when the range is extended to ±45°, as detailed in Tab. 2. This underlines the robustness of our approach. Our approach leverages pixel-wise localization, offering finer granularity compared to the patch-wise localization of SliceMatch[9], CCVPE [28], and the image-level localization of DSM [15]. Although HighlyAcc [14] and HC-Net [26] are pixel-wise localization methods, their homography-based mechanisms fail to incorporate off-road cues. In contrast, our approach effectively utilizes these cues, resulting in enhanced performance. Additionally, our longitudinal estimation significantly surpasses that of BoostAcc [17]. This improvement is likely attributable to our method's strategy of not aggregating information from pixels below, thereby reducing longitudinal ambiguity. In comparison to the baseline method PureACL [24], our approach achieves significant enhancements in both orientation and spatial accuracy, especially in mean error reduction. Specifically, we observe a $91^{+}\%$ reduction in spatial mean error and a $52^{+}\%$ reduction in orientation mean error. The difference between mean and median errors for PureACL indicates potential convergence issues in high-noise environments. By integrating the T2GA and CycDA modules, we bridge the domain gap and ensure convergence even under challenging conditions.

5.3. Challenge Initial Pose and Stringent Metrics

To rigorously evaluate our method, we adopt the stringent metrics of PureACL [24], which are more demanding than those used by HighlyAcc[14]. We benchmark our method against baseline methods SIBCL [25] and PureACL [24], as well as SOTA methods CCVPE [28] and BoostAcc [17]. The results in Tab. 2 reaffirm the advantages of our

Table 2. Comparison with Initial Noise Conditions (±45°,±20m).For the KITTI-CVL dataset, evaluations are conducted in 'K-Same' and 'K-Cross' areas. On the Ford-CVL dataset, the 'Log4' trajectory, as used in SIBCL [25], is chosen for its optimal satellite view alignment. Additional log evaluations are detailed in the supplementary material. 'F-1C' and 'F-4C' represent assessments on the Ford-CVL dataset using either a single front-facing camera or four surrounding cameras. Note: ⋆ indicates that SIBCL [25] is a hybrid LiDAR-visual method.

		Lateral(m)↓		Lateral(%)↑			Longit(m)↓		Longit(%)↑			Orient(°)↓		Orient(%)↑		
	Model	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°
K-Same	CCVPE [28]	2.32	1.22	11.05	22.05	42.09	6.49	3.35	7.45	13.36	24.97	3.43	1.53	34.66	60.72	84.03
	BoostAcc[17]	1.19	0.63	21.92	41.51	67.72	10.01	5.42	5.17	9.38	16.54	3.88	2.98	18.45	35.25	62.20
	⋆ SIBCL[25]	1.23	0.70	17.01	33.34	61.88	3.14	0.86	20.68	31.40	55.58	8.90	1.69	35.50	55.04	71.06
	PureACL[24]	2.94	0.23	58.75	66.34	70.60	3.15	0.37	45.65	68.23	74.87	8.13	2.80	37.99	52.57	64.13
	Ours	**0.17**	**0.14**	**76.20**	**97.73**	**99.97**	**0.08**	**0.07**	**98.18**	**99.89**	**99.97**	**2.19**	**1.07**	**48.97**	**66.29**	**90.54**
K-Cross	CCVPE [28]	4.56	2.79	4.96	10.05	38.12	12.06	8.16	3.13	6.14	12.45	19.46	15.39	3.38	10.24	17.41
	BoostAcc[17]	2.65	0.93	14.88	29.20	52.64	10.46	8.92	3.38	5.56	11.73	5.61	4.26	12.46	24.87	47.26
	⋆ SIBCL[25]	2.72	0.71	16.90	32.62	58.24	5.48	1.05	19.58	30.74	49.25	9.67	1.96	25.02	50.18	60.16
	PureACL [24]	4.24	0.25	50.25	62.63	64.56	3.89	0.44	7.41	62.04	64.23	10.96	3.16	19.95	36.25	56.49
	Ours	**0.17**	**0.15**	**75.96**	**97.47**	**99.97**	**0.09**	**0.07**	**97.49**	**99.87**	**99.97**	**2.88**	**1.44**	**36.10**	**63.08**	**86.18**
F-1C	CCVPE [28]	4.62	2.45	6.48	12.54	24.46	10.21	8.41	3.22	8.24	13.53	20.48	15.07	3.14	6.14	16.14
	BoostAcc[17]	2.73	1.47	9.05	18.18	35.59	10.51	6.14	4.21	8.13	15.56	6.82	4.79	11.61	22.64	42.63
	⋆ SIBCL[25]	2.59	0.71	20.92	41.53	60.43	5.38	1.18	12.59	23.99	43.87	6.34	1.57	35.21	58.41	70.65
	PureACL[24]	2.94	1.54	11.34	21.14	37.76	4.75	1.74	10.46	20.07	37.64	7.38	3.33	15.12	30.42	57.01
	Ours	**0.49**	**0.41**	**31.77**	**59.17**	**89.23**	**0.36**	**0.32**	**38.43**	**75.72**	**97.46**	**2.19**	**1.08**	**47.31**	**72.71**	**90.10**
F-4C	CVGL[4]	1.28	0.50	21.81	50.05	86.84	1.53	0.70	19.36	37.26	65.51	-	-	-	-	-
	PureACL[24]	1.69	0.67	21.07	39.53	65.27	3.38	1.18	12.59	23.99	43.86	3.07	1.19	43.84	73.27	82.41
	Ours	**0.13**	**0.10**	**88.25**	**98.53**	**100.00**	**0.20**	**0.18**	**68.67**	**96.68**	**100.00**	**2.00**	**1.49**	**35.72**	**62.70**	**91.83**



(a) Location

(b) Orientation

Figure 6. Comparison of the baseline method PureACL and our proposed method under varying initial noise ranges in the 'cross' area of the KITTI-CVL dataset. The blue bar indicates the median value, while the red bar shows the difference between the median and mean values. Notably, our method demonstrates robust convergence, outperforming PureACL across all noise ranges, especially in larger noise scenarios where PureACL shows significant discrepancies between median and mean values.

method, particularly in terms of reduced mean error and improved longitudinal estimation. Notably, under conditions of minimal initial pose noise (±15° and ±5m), PureACL [24] outperforms SIBCL [25]. However, when faced with more challenging conditions (±45° and ±20m), PureACL [24] encounters convergence challenges, unlike our method, which demonstrates consistent superiority. This highlights the effectiveness of our proposed modifications in feature alignment. Additionally, we assess our method's convergence range against PureACL [24] in the 'Cross' area of the KITTI-CVL dataset, as depicted in Fig. 6. The red bar in the figure highlights a significant reduction in the gap between median and mean errors. This observation confirms our method's enhanced ability to converge effectively across a wider range of noise levels.

5.4. Results with Continual Pose Estimation

To address GPS signal loss, we adopt an accumulated pose estimation strategy that leverages initial poses based on the vehicle's previous pose estimates [7]. We use the model trained with initial noise allowances of ±45° and ±20m. Our method's robustness is evidenced by the continuous running distance percentage presented in Tab. 3, showing successful pose chaining throughout all evaluate routes. In contrast, BoostAcc [17] and the single-camera PureACL [24] exhibit significant drift, leading to errors beyond the satellite map's coverage and causing evaluations to cease after covering less than 11% of the distance. Performance comparisons on the KITTI-CVL dataset are shown in Fig. 7 [8]. Our approach consistently achieves accurate pose estimation or experiences only minor drifts in challenging scenarios characterized by limited localization cues or severe occlusions. In contrast, BoostAcc [17] exhibits significant drift when a moving vehicle passes by, likely due to incorrect query data from projected vehicle pixels. Furthermore, PureACL [24] appears to struggle with incorrect orientation, resulting in reversed pose estimations. These out-

[7] This pose estimation is purely frame-based and does not involve any sequence-based filtering techniques.

[8] Performance comparisons on the Ford-CVL dataset is shown in Fig. 1 and Fig. 2 of supplementary material.

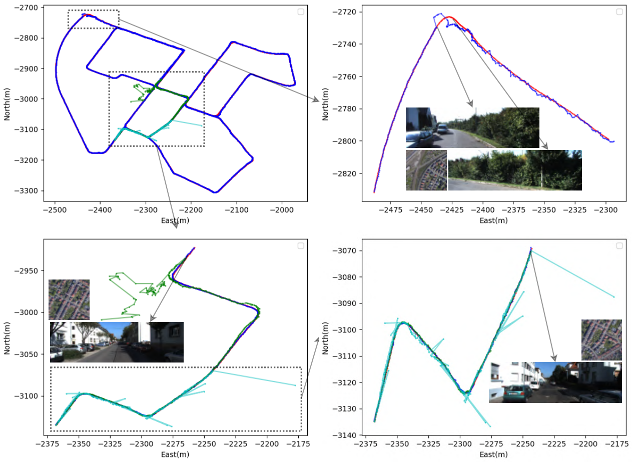


Figure 7. Accumulated pose estimation performance on a KITTI-CVL dataset trajectory. (Top-left) Predicted poses by our method (blue) closely match the ground truth (red) over the 3676-meter route, outperforming PureACL (green) and BoostAcc (cyan), which exhibit severe drift. (Top-right) Our approach demonstrates rapid recovery upon encountering clear localization cues. (Bottom-left) (Bottom-right) BoostAcc demonstrates substantial drift when encountering moving vehicles.

comes highlight the importance of robustness in real-world applications, as drifts that extend beyond the coverage of satellite imagery can compromise all further pose estimation processes.

Table 3. Comparison on Accumulated Pose Estimation.

Model	Running percentage(%)↑		
	K-Cross	**F-1C**	**F-4C**
BoostAcc[17]	4.88	4.89	-
PureACL[24]	10.31	5.01	**100.00**
Ours	**100.00**	**100.00**	**100.00**

- indicates that BoostAcc does not support 4-camera setting.

6. Ablation Study

To evaluate the effectiveness of our proposed components, we conduct ablation studies in the 'cross' area of the KITTI-CVL dataset. Our method's performance was compared across various configurations, including the presence and absence of CycDA and Equidistant Re-projection (ERP) Loss functions, and the T2GA module. The results, as detailed in Tab. 4, underscore the significant role each component plays in improving our method's performance.

The T2GA module aggregates information from top to ground based on the assumption that the ground and satellite views are orthogonal. However, this assumption might not be accurate in scenarios involving uphill/downhill or turns. To test our method's robustness under these conditions, we introduced affine warping to the ground view images with a shear range of $\pm 15^{\circ}$. The outcomes, labeled 'w/ Affine' in Tab. 4, indicate our method's efficacy in handling such scenarios. This success is primarily due to two factors: (1) the use of a U-Net for feature extraction allows each pixel in the feature map to encapsulate information from its surrounding area, enabling the handling of a degree of angular variation. (2) the satellite view from Google, which is not strictly orthogonal to the ground view, suggests that our method is already adapted to these types of challenges.

Table 4. Ablation study on the KITTI-CVL dataset 'Cross' area with initial noise ($\pm 45^{\circ}$,$\pm$20m).

			Lateral(m)↓		Longit(m)↓		Orient(°)↓	
T2GA	CycDA	ERP	Mean	Median	Mean	Median	Mean	Median
			4.24	0.25	3.89	0.44	10.96	3.16
✓			1.29	0.19	1.27	0.17	4.11	2.13
✓	✓		0.21	0.18	0.12	0.10	3.38	1.89
✓	✓	✓	**0.17**	**0.15**	**0.09**	**0.07**	**2.88**	**1.44**
w/ Affine			0.18	0.15	0.09	0.07	2.88	1.52
night			0.39	0.27	0.39	0.30	6.64	2.55

We further evaluate our method on an artificial dataset derived from the KITTI test sets, created using the cross-domain deep network proposed by Arruda et al. [2]. The outcomes, labeled as 'night' in Tab. 4 show our method's stable performance in nighttime scenarios, highlighting its resilience to variations in lighting and time of day. For further details, please refer to Supplementary Sec. C.

7. Conclusion

We have presented a novel top-to-ground feature aggregation for enhancing cross-view image-based geo-localization. Our method overcomes limitations by incorporating aerial perspectives and utilizing a cycle domain adaptation loss for consistent feature extraction despite visual disparities. The introduction of the equidistant re-projection loss balances keypoint impact, promoting wider distribution and thus enhancing orientation accuracy. Our method excels in vehicle pose estimation across challenging scenarios, achieving the lowest translation errors in KITTI and Ford datasets, and minimal orientation error with less accurate initial poses. Crucially, by relying solely on the initial vehicle pose at the start, our method successfully completes routes via continuous pose estimation. This paves the way for real-world applications like autonomous driving and outdoor robotics. Future work will focus on integrating this method into SLAM systems to remove loop closure dependency and facilitate high-definition map generation.

Acknowledgements. This project is supported by CSIRO Julius Career Award, CSIRO Early Research Career Fellowship, CSIRO Future Digital Manufacturing Fund Postdoc Fellowship, CSIRO Data61 PhD Scholarship to SW, and partially by Australian Research Council Discovery Grant (DP220100800) to HL.

References

[1] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James McBride. Ford multi-AV seasonal dataset. *The International Journal of Robotics Research*, 39(12):1367–1376, 2020. 5

[2] 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. 8

[3] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 5

[4] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21621–21631, 2023. 1, 2, 7

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5

[6] Google. Maps static api, 2023. 2023. 1

[7] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 1, 2

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[9] Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17225–17234, 2023. 2, 5, 6

[10] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[11] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977*, pages 105–116. Springer, 2006. 3

[12] 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. 3

[13] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. In *CVPR*, 2023. 2

[14] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 5, 6

[15] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 1, 2, 6

[16] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11990–11997, 2020. 2

[17] 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, 3, 6, 7, 8

[18] Shinya Sumikura, Mikiya Shibuya, and Ken Sakurada. Openvslam: A versatile visual slam framework. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2292–2295, 2019. 5

[19] Xiaoyang Tian, Jie Shao, Deqiang Ouyang, and Heng Tao Shen. Uav-satellite view synthesis for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4804–4815, 2021. 1, 2

[20] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 1, 2

[21] Frank Van Diggelen and Per Enge. The world's first gps mooc and worldwide laboratory using smartphones. In *Proceedings of the 28th international technical meeting of the satellite division of the institute of navigation (ION GNSS+ 2015)*, pages 361–369, 2015. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[23] Shan Wang, Chuong Nguyen, Jiawei Liu, Kaihao Zhang, Wenhan Luo, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, and Hongdong Li. Homography guided temporal fusion for road line and marking segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1075–1085, 2023. 1

[24] Shan Wang, Yanhao Zhang, Akhil Perincherry, Ankit Vora, and Hongdong Li. View consistent purification for accurate cross-view localization, 2023. 1, 2, 3, 5, 6, 7, 8

[25] 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, 2, 5, 6, 7

[26] Xiaolong Wang, Runsen Xu, Zuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using

a correlation-aware homography estimator. *arXiv preprint arXiv:2308.16906*, 2023. 1, 2, 3, 6

[27] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 90–106. Springer, 2022. 2

[28] Zimin Xia, Olaf Booij, and Julian FP Kooij. Convolutional cross-view pose estimation. *arXiv preprint arXiv:2303.05915*, 2023. 2, 5, 6, 7

[29] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh. Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE Transactions on Multimedia*, 25:2176–2188, 2023. 1, 2

[30] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 1, 2

[31] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 1, 2