VIRD: View-Invariant Representation through Dual-Axis Transformation for Cross-View Pose Estimation

Supplementary Material

Table of Contents

Unless otherwise specified, all experiments are conducted on the KITTI [5] dataset under the cross-area setting.

A. Context-Enhanced Positional Attention

A.1. Pseudocode

Algorithm 1 Pseudocode for CEPA

Input:
  Query positional encoding P_a # (H_Q, d_p)
  Key positional encoding P_v # (H_K, d_p)
  Ground attention weights A_g # (H_Q, H_K)
  Ground feature F_g # (B, C, H_K, W_g)
Output:
  Positional attention weights A_v # (H_Q, H_K)
  Context-enhanced attention weights $A_{g'}$ # (B, H_Q, H_K, W_g)

function CAL_POSITIONAL_ATTN_WEIGHTS(P_a, P_v)
  $Q \leftarrow$ linear_projection(P_a)
  $K \leftarrow$ linear_projection(P_v)
  $A_v \leftarrow (QK^\top)/\sqrt{d_k}$
  $A_v \leftarrow$ softmax(A_v, dim $= -1$)
  return A_v
end function

function CONTEXT_ENHANCEMENT(A_g, F_g)
  $A'_g \leftarrow$ reshape(A_g, $1, H_Q, 1, H_K, 1$)
  $A'_g \leftarrow$ repeat(A'_g, $B, 1, 1, 1, W_g$) # $(B, H_Q, 1, H_K, W_g)$

  $F'_g \leftarrow$ reshape(F_g, $B, 1, C, H_K, W_g$)
  $F'_g \leftarrow$ repeat(F'_g, $1, H_Q, 1, 1, 1$) # (B, H_Q, C, H_K, W_g)

  feat $\leftarrow$ concat(A'_g, F'_g, dim $= 2$) # $(B, H_Q, C{+}1, H_K, W_g)$
  feat $\leftarrow$ reshape(feat, $B{\cdot}H_Q$, $C{+}1$, H_K, W_g)
  feat $\leftarrow$ conv2d(feat) # $(B{\cdot}H_Q, 1, H_K, W_g)$
  feat $\leftarrow$ reshape(feat, B, H_Q, H_K, W_g)
  $\hat{A} \leftarrow$ softmax(feat, dim $= -2$)
  $A_{g'} \leftarrow \hat{A} + A'_g$ # after broadcasting
  return $A_{g'}$
end function

Algorithm 2 Pseudocode for vertical feature transformation

Input:
  Shared virtual positional encoding P_a # (H_Q, d_p)
  Ground positional encoding P_g # (H_K, d_p)
  Satellite positional encoding P_{s2p} # (H_K, d_p)
  Ground feature F_g # (B, C, H_K, W_g)
  Polar-transformed satellite feature F_{s2p} # (B, C, H_K, W_s)
Output:
  Vertically-transformed ground feature $F_{g'}$ # (B, C, H_Q, W_g)
  Vertically-transformed satellite feature $F_{s2p'}$ # (B, C, H_Q, W_s)

function VERTICAL_TRANSFORM($P_a, P_g, P_{s2p}, F_g, F_{s2p}$)
  $A_g \leftarrow$ CAL_POSITIONAL_ATTN_WEIGHTS(P_a, P_g)
  $A_{s2p} \leftarrow$ CAL_POSITIONAL_ATTN_WEIGHTS(P_a, P_{s2p})
  $A_{g'} \leftarrow$ CONTEXT_ENHANCEMENT(A_g, F_g)
  $F_{g'}[b, c, h_q, w] \leftarrow \sum_k A_{g'}[b, h_q, h_k, w]\ F_g[b, c, h_k, w]$
  $F_{s2p'}[b, c, h_q, w] \leftarrow \sum_k A_{s2p}[h_q, h_k]\ F_{s2p}[b, c, h_k, w]$
  return $F_{g'}, F_{s2p'}$
end function

A.2. CEPA vs. content-based attention

Fig. A compares the proposed context-enhanced positional attention (CEPA) with a content-based attention (CBA) baseline. CBA follows the conventional formulation in which attention weights are computed directly from input content, such as image features [17]. We implement CBA using the self- and cross-attention layers from SliceMatch [11] and apply them with a polar transformation (Polar + CBA). CBA yields a moderate gain in the same-area setting, where training and test scenes share similar visual patterns. However, its performance degrades in the cross-area setting. Because CBA derives attention weights from image content, it overfits to scene-specific semantics and fails to capture the spatial correspondence across unseen environments. In contrast, CEPA consistently achieves superior performance in both same- and cross-area evaluations. It first derives attention weights from positional cues and then refines them using ground context, allowing the model to more effectively focus on spatial correspondence rather than appearance. These results demonstrate that explicitly modeling positional correspondence is crucial for robust cross-view pose estimation.

A.3. Positional encoding types

Tab. A compares different types of positional encodings, including sinusoidal, sinusoidal-learnable, learnable, and 2D sinusoidal variants. The sinusoidal encoding is fixed and follows the formulation of [26]. The sinusoidal-learnable variant applies a nonlinear projection to the fixed sinusoidal encoding. In contrast, the learnable encoding is initialized as a fully trainable tensor without relying on any predefined

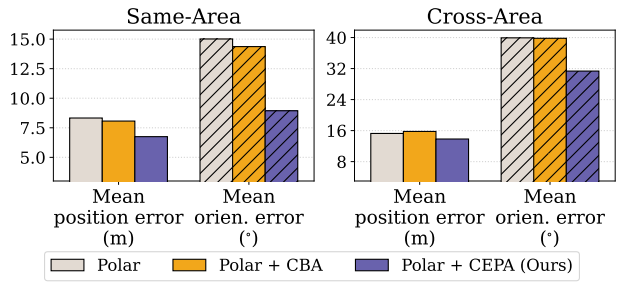


Figure A. Comparison of content-based attention (CBA) and context-enhanced positional attention (CEPA) on KITTI [5].

Table A. Positional attention performance with various positional encoding variants on KITTI [5].

	Pos. error (m) ↓		Orien. error (°) ↓	
	Mean	Med.	Mean	Med.
Sinusoidal pe	**13.92**	**9.76**	35.17	3.44
Sinusoidal learnable pe	14.28	10.63	33.82	**3.41**
Learnable pe	15.23	12.06	36.23	4.04
2D sinusoidal pe	14.68	10.31	**33.57**	3.57

positional pattern. The 2D sinusoidal variant extends positional information to both horizontal and vertical directions in the ground view.

Among these variants, the fixed 1D sinusoidal encoding (Polar + PA in the main paper's Tab. 3) achieves the best overall performance, showing the lowest position error and competitive orientation accuracy. These results indicate that vertical transformations across views can be learned more reliably when positional encodings are fixed and 1D. Notably, the comparison with the 2D sinusoidal variant suggests that encoding both horizontal and vertical positional information may introduce ambiguity in establishing spatial correspondence. This may occur because the standard positional attention mechanism relies solely on positional information, which might not adequately account for the complexity of vertical structures in ground-view images. In contrast, assuming a consistent vertical transformation along the horizontal direction provides a more stable inductive bias that facilitates reliable spatial correspondence. This finding further highlights the strength of the proposed CEPA, which uses fixed 1D positional encoding to model shared vertical transformation while adaptively capturing horizontal variations in the ground view through context enhancement.

B. View-Reconstruction Loss

B.1. Effect of view-reconstruction on CEPA

This section analyzes how the view-reconstruction loss influences the learning of CEPA. During training, the view-reconstruction loss encourages the ground and satellite descriptors to reconstruct both their original and cross views. This supervision enables the descriptors to embed structural cues more effectively, not only from regions with small viewpoint gaps, such as road surfaces, but also from structures with large viewpoint gaps, including tall buildings and other vertically dominant elements. By promoting the preservation of meaningful information even in these challenging regions, the view-reconstruction loss allows CEPA to attend more effectively to vertical structures and ultimately learn more stable vertical transformations between the cross views. As shown in Fig. B, the visualization of the context-enhanced ground attention weights highlights this effect. In example (a), CEPA with the view-reconstruction loss exhibits stronger activations around roof regions, indicating improved vertical correspondence. In example (b), the model successfully captures the overpass structure that is missed when the view-reconstruction loss is removed.

Table B. Comparison of different reconstruction loss types on KITTI [5].

	Pos. error (m) ↓		Orien. error (°) ↓	
	Mean	Med.	Mean	Med.
ℓ_1	**12.72**	**7.90**	**21.91**	**3.05**
ℓ_2	13.14	8.30	25.74	3.31
SSIM	12.92	8.00	25.23	3.34
Percep	14.00	9.78	29.55	3.31

B.2. Reconstruction loss types

Several loss functions for view reconstruction are evaluated, including ℓ_1, ℓ_2, SSIM, and perceptual loss, as summarized in Tab. B. Among them, the ℓ_1 loss achieves the best performance, yielding a mean position error of 12.72 m and a mean orientation error of 21.91°.

The goal of view reconstruction in our framework is not to generate perceptually realistic images, but to provide stable structural supervision that promotes view-invariant descriptors. ℓ_2 loss is less suitable because it strongly penalizes large pixel differences that naturally arise between ground and satellite views, leading to unstable gradients. SSIM focuses on luminance and contrast consistency, which are not reliably shared across cross views, thereby weakening the geometric cues needed for spatial correspondence. Perceptual loss is also ineffective: it compares high-level semantic features extracted by networks trained on perspective natural images, but such semantics differ substantially between ground and satellite views, producing inconsistent gradients that degrade geometric structure. In contrast, the ℓ_1 loss provides uniformly weighted, geometry-preserving supervision that avoids overfitting to appearance or semantics. This property makes the ℓ_1 loss the most stable and effective choice for enforcing view invariance in cross-view matching.

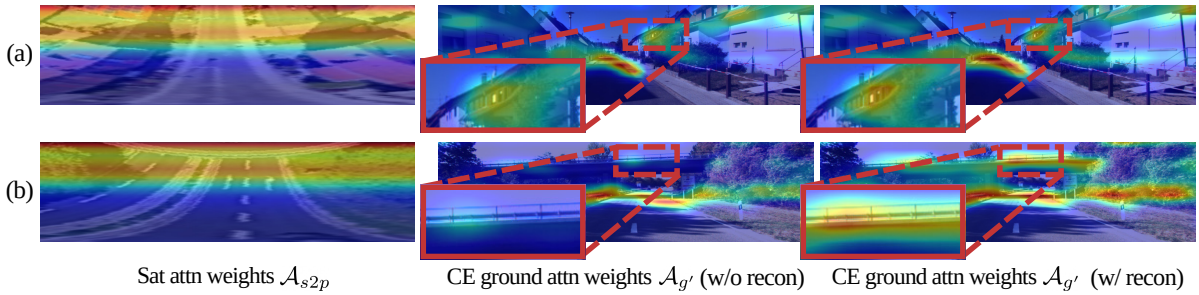


Figure B. Effect of the view-reconstruction loss on CEPA. The first row presents example (a), and the second row presents example (b). For each example, the satellite attention $\mathcal{A}_{s2p}$ and the context-enhanced ground attention $\mathcal{A}_{g'}$ are visualized at the same vertical position in the shared virtual positional encoding space. The view-reconstruction loss stabilizes the learning of vertical transformations in CEPA, yielding stronger roof activations in example (a) and successful detection of overpass structures in example (b). Regions with stronger activations are highlighted in red.

C. Matching and Regression Modules

C.1. Shifting and cropping of satellite descriptor

Given a candidate pose $\mathbf{p}_c = (x_c, y_c, \theta_c)$, a satellite descriptor $D_{s2p}^{(x_c,y_c)} \in \mathbb{R}^{K_s}$ is constructed via dual-axis transformation and vertical directional encoding. This descriptor is initially aligned such that its central vertical line corresponds to the east, i.e., $\theta = 0$. To align it with the orientation θ_c, the descriptor is cyclically shifted along the horizontal axis by $\frac{\theta_c}{2\pi} \cdot W_s$, following prior work [35]. After alignment, the descriptor is center-cropped to match the size of the ground descriptor $D_g \in \mathbb{R}^{K_g}$, yielding the orientation-aligned satellite descriptor $D_{s2p}^{\mathbf{p}_c} \in \mathbb{R}^{K_g}$. Similarly, the polar-transformed satellite image $I_{s2p}^{\mathbf{p}_c}$, which is used in the view-reconstruction loss, is shifted and cropped in the same manner to ensure spatial alignment with I_g.

C.2. Regression network architecture

The regression module first reshapes the 1D descriptor difference into a 2D representation with channel and width dimensions, then passes it through a series of 1D convolutional layers to obtain a high-dimensional embedding. The resulting feature is flattened and concatenated with the coarse pose. MLPs are applied to predict the pose residual $\Delta\mathbf{p} = (\Delta x, \Delta y, \Delta\theta)$ relative to the ground truth. The predicted output is scaled to operate within a predefined search range.

C.3. Training details

Matching loss. To train discriminative descriptors, we adopt the InfoNCE loss [14], encouraging high similarity at the ground-truth pose $\mathbf{p}^*$ and penalizing n-th non-matching pose $\mathbf{p}_n$, following [11, 35]:

$$\mathcal{L}_{\text{match}} = -\log\left(\frac{\exp(S^{\mathbf{p}^*}/\tau)}{\sum_{n=1}^{N}\exp(S^{\mathbf{p}_n}/\tau) + \exp(S^{\mathbf{p}^*}/\tau)}\right), \tag{7}$$

where $S^{\mathbf{p}^*}$ and $S^{\mathbf{p}_n}$ are similarity scores at the ground-truth pose and the n-th non-matching pose among N samples, respectively. τ is a temperature scaling parameter.

Regression loss. The regression module is trained using the ℓ_2-loss computed over pose residuals:

$$\mathcal{L}_{\text{reg}} = \beta(\|\Delta x - \Delta x^*\|_2 + \|\Delta y - \Delta y^*\|_2 + \|\Delta\theta - \Delta\theta^*\|_2), \tag{8}$$

where $(\Delta x^*, \Delta y^*, \Delta\theta^*)$ denote the ground-truth pose residuals, and β is a weighting coefficient.

Regression training. To train the regression module, N_r coarse poses are randomly sampled around the ground-truth pose within predefined spatial and angular ranges, where reliable refinement is feasible. For each sampled pose $\mathbf{p}_r$, the satellite descriptor $D_{s2p}^{\mathbf{p}_r}$ is computed using the same shift-and-crop mechanism described in Sec. C.1. The regression loss for each sampled pose is computed as described in Eq. (8), and the final loss is obtained by averaging over all N_r samples.

D. Additional Implementation Details

We provide additional implementation details in Sec. 4.3. The spatial size of I_g is 256×1024 for KITTI and 320×640 for VIGOR, and I_s is 512×512 for both datasets. The temperature parameter τ is set to 0.05. The loss coefficients α_1, α_2, and β are set to 1, 10, and 5, respectively. The polar transform is configured with $r_{\min}$ and $r_{\max}$ set to 6 m and 40 m for KITTI [5], and to 0 m and 30 m for VIGOR [38], respectively. For the CEPA module, the shared virtual axis height H_Q is set equal to the ground feature height H, which is 16 and 20 for KITTI and VIGOR, respectively. To improve computational efficiency, the number of channels is reduced by a factor of four after feature extraction with the VGG16 backbone. The feature extractor weights are shared when using the VGG16 backbone; in all other

Table C. Impact of cylindrical projection on pinhole camera on KITTI [5].

Cylindrical projection	Pos. error (m) ↓ Mean	Med.	Orien. error (°) ↓ Mean	Med.
×	15.26	11.75	39.92	**4.00**
✓	**15.10**	**11.62**	**38.97**	4.13

Table D. Ablation study of the temperature parameter on KITTI [5].

τ	Pos. error (m) ↓ Mean	Med.	Orien. error (°) ↓ Mean	Med.
0.1	16.09	13.21	**38.54**	4.34
0.05	**15.26**	**11.75**	39.92	**4.00**
0.01	17.18	14.93	53.59	6.17

settings, weights are not shared. Zero-padding is used as the default setting. For panorama images, circular padding is applied along the horizontal axis. For regression, the search range is limited to ± 4 m, ± 4 m, and $\pm 3.6°$ in x, y, and θ dimensions, respectively.

E. Applicability to Pinhole Camera Models

Following prior work [35], we assume a cylindrical projection for all cameras. This assumption is not strictly valid for pinhole cameras, meaning that the polar-transformed satellite image does not perfectly correspond to the horizontal axis of a pinhole ground image. Despite this mismatch, the polar transformation remains effective for reducing the viewpoint gap. As shown in Tab. C, models trained with raw pinhole images and those trained with a cylindrical projection applied to the same images yield nearly identical performance. This result indicates that the projection mismatch has negligible influence on accuracy, demonstrating that VIRD can be applied to pinhole cameras without difficulty.

F. Additional Ablation Studies

F.1. Hyperparameters

Temperature The temperature parameter τ is evaluated using a baseline model that includes only the polar transformation, excluding the proposed attention and reconstruction modules. Tab. D shows that the best performance is achieved at 0.05.

Loss coefficients Tab. E presents the results of ablation experiments on the loss coefficients. The best performance is achieved when the loss coefficients for original-view reconstruction (α_1), cross-view reconstruction (α_2), and regression (β) are set to 1, 10, and 5, respectively. The matching loss coefficient is set to its default value of 1, and thus no ablation was performed for this parameter.

Table E. Ablation study of the loss coefficients on KITTI [5].

α_1	α_2	β	Pos. error (m) ↓ Mean	Med.	Orien. error (°) ↓ Mean	Med.
0.1	0	0	13.43	8.54	30.08	**3.30**
1	0	0	**13.34**	**8.29**	28.26	3.31
5	0	0	13.90	9.04	**27.01**	3.34
1	5	0	**12.61**	**7.64**	25.72	3.16
1	10	0	12.72	7.90	**21.91**	**3.05**
1	20	0	13.25	8.39	24.13	3.54
1	10	1	12.45	7.48	**23.59**	2.47
1	10	5	**12.30**	**7.05**	25.10	**2.22**

Table F. Ablation study of the regression search range on KITTI [5].

Range	Area	Lat. (%) ↑ R@1m	Lon. (%) ↑ R@1m	Orien. (%) ↑ R@1°
± 2 m, ± 2 m, $\pm 1.8°$	Same	79.30	29.21	44.16
± 4 m, ± 4 m, $\pm 3.6°$	Same	**79.46**	**31.65**	49.32
± 6 m, ± 6 m, $\pm 4.8°$	Same	74.77	28.49	**51.63**
± 2 m, ± 2 m, $\pm 1.8°$	Cross	**44.37**	**12.98**	26.81
± 4 m, ± 4 m, $\pm 3.6°$	Cross	43.61	12.88	**27.65**
± 6 m, ± 6 m, $\pm 4.8°$	Cross	40.20	12.11	27.06

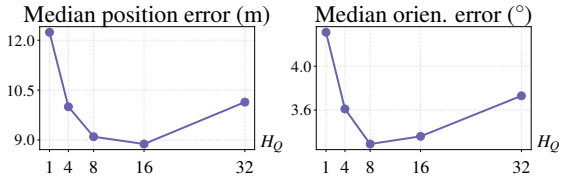


Figure C. Ablation study of H_Q on KITTI [5].

Regression search range Tab. F reports the effect of varying the search range for the regression module. The model achieves high average recall accuracy when the search range is limited to ± 4 m in translation and $\pm 3.6°$ in orientation. This setting performs comparably to the tighter configuration of ± 2 m and $\pm 1.8°$, while significantly outperforming broader ranges such as ± 6 m. These results demonstrate that the optimal search range for the regression module lies within a moderate range, where it can more precisely focus on local feature differences and more effectively correct residual pose errors.

F.2. Impact of shared virtual axis height

As shown in Fig. C, we conduct an ablation study on the height of the shared virtual axis H_Q, which defines the resolution of the shared coordinate system in the CEPA module. Performance peaks around $H_Q = 16$, demonstrating that this resolution is sufficient to capture vertical correspondences between cross views. Notably, performance drops markedly at $H_Q = 1$, where vertical information is simply collapsed into a single dimension without establishing vertical correspondence. This result supports the necessity of the

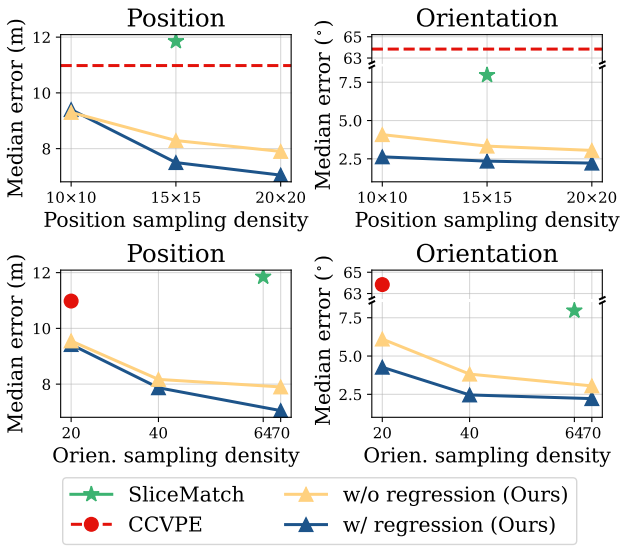


Figure D. Ablation study of candidate pose sampling density, compared with SliceMatch [11] and CCVPE [35]. For the proposed method, the number of orientation candidates is fixed to 70 during position sampling, and the position grid size is fixed to 20×20 during orientation sampling. SliceMatch performs matching only on 15×15×64 candidate poses. For CCVPE, position prediction is performed at the original resolution, while orientation prediction first matches 20 candidates and then performs regression. Because CCVPE is independent of the number of position samples, its performance is shown as a horizontal line.

shared virtual axis for resolving vertical misalignment. On the other hand, increasing H_Q to 32 leads to degraded performance, suggesting that excessively high resolution may introduce redundancy that hinders stable learning.

F.3. Effect of candidate pose sampling density

An ablation study is conducted to examine the impact of candidate pose sampling density, with the analysis separated into position and orientation sampling. For position sampling, three spatial resolutions are compared: 10×10, 15×15, and 20×20. For orientation sampling, four candidate counts are evaluated: 20, 40, 64, and 70. These sampling configurations are selected with reference to prior works [11, 35].

Fig. D shows that increasing the number of candidate poses leads to reduced median errors in both position and orientation. This trend highlights the benefit of finer sampling resolution. Although denser sampling improves accuracy, it also increases computational cost. Sec. H.2 examines this trade-off in more detail. We also compare two global descriptor-based approaches: SliceMatch [11] and CCVPE [35]. The proposed method consistently improves over all baselines across the tested sampling densities, demonstrating the superiority of our approach.

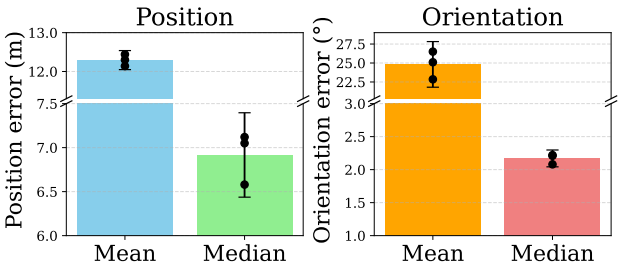


Figure E. Error statistics across five random seeds for the proposed method. Error bars indicate $\pm 2\sigma$ standard deviations, computed from five independent runs.

G. Error Statistics across Random Seeds

To assess the robustness of the proposed method, performance variation is reported across the five independent runs using different random seeds. Fig. E visualizes the mean and median errors for both position and orientation, providing a statistical perspective on the results presented in Tab. 1. Error bars represent $\pm 2\sigma$ standard deviations, computed as the sample standard deviation across five independent runs. This captures variability due to random initialization and training dynamics. While the normality assumption is not formally tested due to the limited number of samples, the visualization provides a practical view of the method's stability across repeated runs.

H. Computational Resources and Runtime

H.1. Training and inference

We report the computational cost during both training and inference. All experiments were conducted on a local workstation equipped with an RTX A5000 GPU with 24 GB of memory. A batch size of 4 was used during training and 1 during inference. During training, the peak GPU memory usage reached approximately 24 GB on both the KITTI and VIGOR datasets. Using the VGG16 backbone, training required about 15 hours on KITTI and 38 hours on VIGOR.

Tab. G summarizes GPU memory usage and runtime during inference on the KITTI dataset. With the VGG16 backbone and a sampling density of $20 \times 20 \times 70$, VIRD requires 9.0 GB of GPU memory and achieves 12 frames per second (FPS). Using EfficientNet-B0 yields similar memory usage while slightly improving runtime to 13 FPS. Compared with prior methods, the proposed approach offers substantially higher accuracy while maintaining comparable memory usage and fast inference speed. These results demonstrate that VIRD is computationally efficient and practical for real-world deployment.

Table G. Inference memory usage and runtime of the proposed method and prior approaches on the KITTI [5] dataset, including the impact of different sampling densities. Sampling density is defined as the number of candidate poses, computed as the product of position and orientation sampling sizes. All results are measured using the official implementations on an RTX A5000 GPU with 24 GB memory.

Method	Backbone	Sampling density	Med. position error	Med. orientation error	GPU memory usage (GB)	Inference runtime (FPS)
HighlyAccurate [18]	VGG16	-	16.02	89.85	7.1	3
CCVPE [35]	EfficientNet-B0	-	10.98	63.84	5.2	20
DenseFlow [23]	ResNet18	-	18.84	42.04	6.4	9
FG2 [31]	DINOv2	-	11.72	90.42	3.5	6
VIRD (Ours)	VGG16	10×10×70	9.40	2.63	4.8	27
VIRD (Ours)	VGG16	15×15×70	7.50	2.35	6.6	18
VIRD (Ours)	VGG16	20×20×70	7.05	2.22	9.0	12
VIRD (Ours)	EfficientNet-B0	20×20×70	5.41	1.87	9.0	13

H.2. Trade-off between sampling density and efficiency

Tab. G also illustrates how increasing the candidate pose sampling density affects computational cost. Higher sampling densities lead to increased memory usage and longer inference times, reflecting a clear trade-off between accuracy and efficiency. Therefore, users may refer to Tab. G to select a sampling configuration that best meets their performance and runtime constraints.

I. Extra Qualitative Results

I.1. Attention weights visualization

We provide additional qualitative examples of attention weights. Fig. F visualizes $\mathcal{A}_g$, $\mathcal{A}_{g'}$, and $\mathcal{A}_{s2p}$, which are computed by applying context-enhanced positional attention to ground and polar-transformed satellite features. The positional attention weights $\mathcal{A}_g$ and $\mathcal{A}_{s2p}$, which are computed without context enhancement, exhibit consistent patterns along the horizontal direction due to learning a shared vertical transformation. In contrast, the context-enhanced attention $\mathcal{A}_{g'}$ adaptively incorporates contextual information for each horizontal position, resulting in more diverse horizontal attention distributions. Overall, these visualizations indicate that the proposed approach effectively captures meaningful spatial correspondence between ground and satellite views.

I.2. Reconstructed images visualization

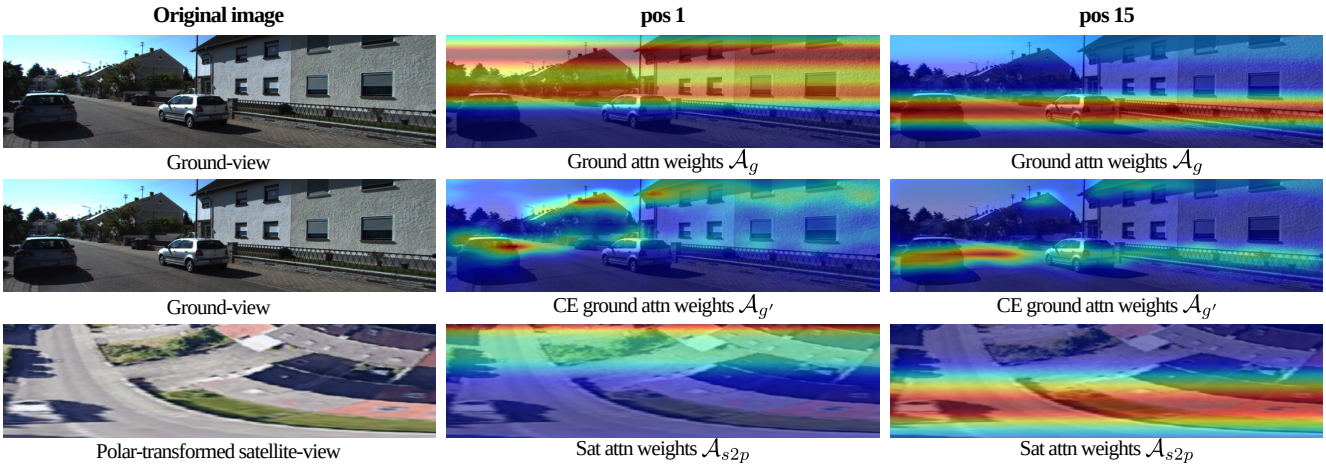
We present additional qualitative examples of reconstructed images. Fig. G visualizes the reconstruction results of the ground and satellite descriptors in both original and cross views. The reconstructed outputs recover road layouts and capture coarse building structures, while naturally suppressing view-specific elements such as vehicles and side roads. These observations indicate that the view-reconstruction loss effectively drives the descriptors toward view-invariant representations that are consistently preserved across ground and satellite views.

J. Societal Impact

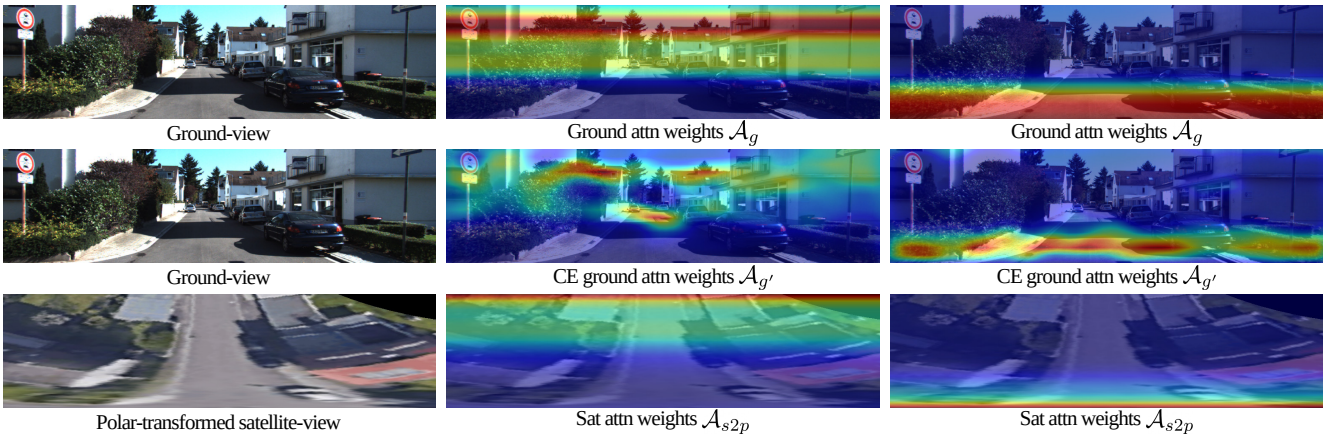
The proposed method advances cross-view localization by enabling accurate ground-to-satellite matching in GNSS-denied environments, such as dense urban areas, where satellite imagery is available. This functionality supports critical applications, including autonomous navigation and disaster response. However, it raises potential privacy concerns and risks of unintended uses, such as surveillance or location-based profiling. These concerns highlight the importance of considering ethical deployment practices and conducting privacy assessments in downstream applications.

K. Limitations

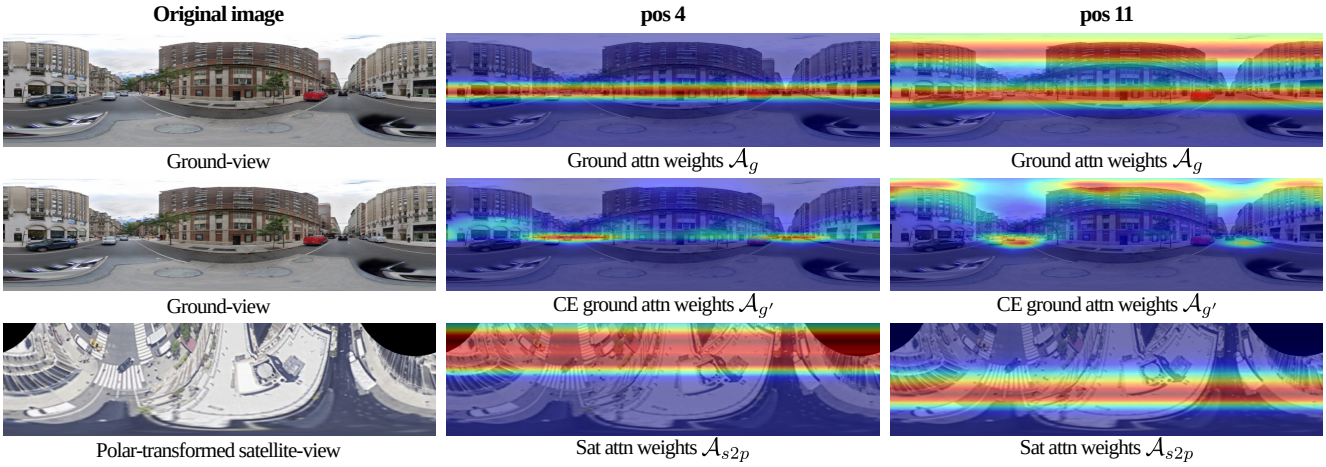
A key limitation of the proposed method is its dependence on the number of candidate poses sampled during inference. Since both model accuracy and computational efficiency are sensitive to the sampling resolution, the method requires manual tuning of the sampling density to meet application-specific constraints, which limits its scalability in practice. In addition, similar to most existing cross-view pose estimation approaches, our method assumes that roll and pitch angles can be neglected. Although this assumption holds in relatively flat urban environments, it may lead to significant errors in challenging terrains such as mountainous regions or areas with strong elevation variations. To address these limitations, future work could focus on reducing the method's reliance on candidate pose sampling resolution and extending the approach toward estimating the full 6-DoF pose without assuming fixed roll and pitch.



(a) Example 1

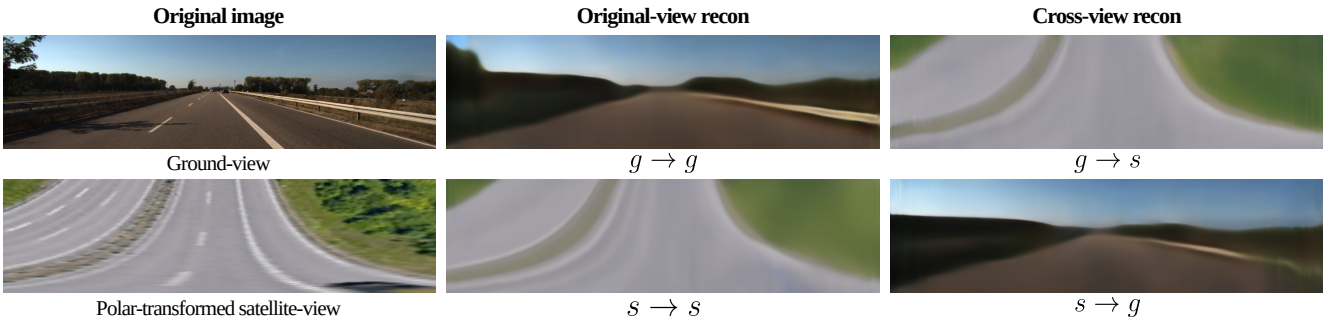


(b) Example 2

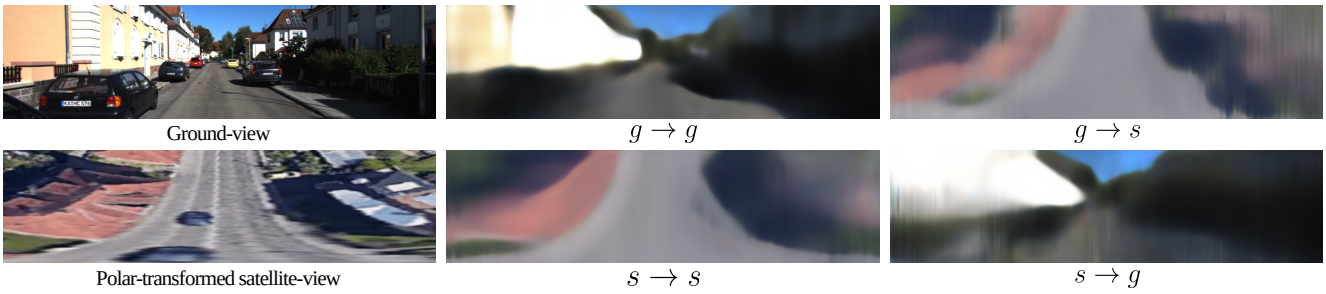


(c) Example 3

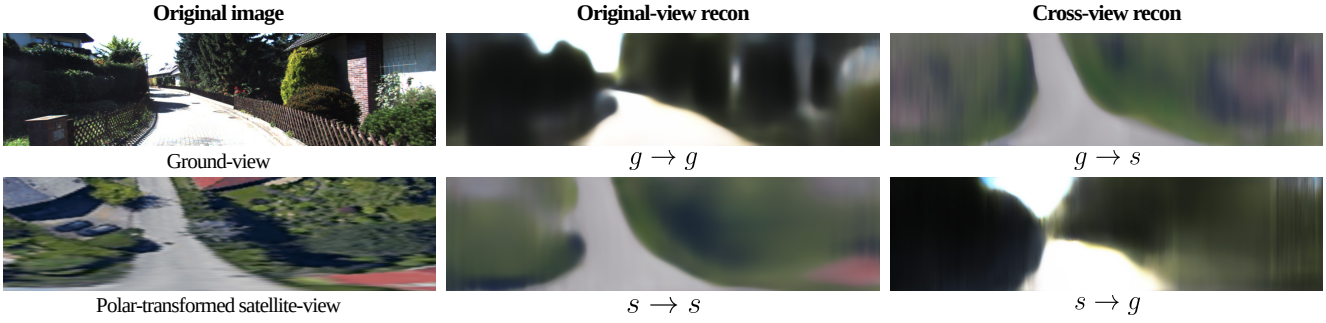
Figure F. Visualization of the attention weights $\mathcal{A}_g$, $\mathcal{A}_{g'}$, and $\mathcal{A}_{s2p}$ at selected positions in the shared virtual positional encoding space. CE and pos denote the context-enhanced attention and the positions in the shared virtual positional encoding space, respectively. Regions with stronger activations are highlighted in red.



(a) Example 1



(b) Example 2



(c) Example 3

Figure G. Visualization of reconstructed images from ground and satellite descriptors in both original and cross views. Each subfigure is denoted as $i \rightarrow j$, where $i, j \in \{s, g\}$ indicate the source and target views, respectively. The pairs $g \rightarrow g$ and $s \rightarrow s$ indicate original-view reconstructions, whereas $g \rightarrow s$ and $s \rightarrow g$ represent cross-view reconstructions.