
Supplementary Material for BevSplat

Anonymous Author(s)

Affiliation

Address

email

1 Robustness to Localization Errors

We evaluate the robustness of our method to varying levels of initial localization error. As shown in Table 1, localization performance improves significantly as the initialization error decreases.

Table 1: Performance comparison under different location error settings on KITTI dataset.

Location Error (m ²)	λ_1	Same Area		Cross Area	
		Mean(m) ↓	Median(m) ↓	Mean(m) ↓	Median(m) ↓
56 × 56	0	5.82	2.85	7.05	3.22
	1	2.87	2.06	6.20	2.51
28 × 28	0	3.27	2.28	3.60	2.47
	1	2.43	1.94	3.31	2.21

2 Ablation Study on Gaussian Primitive Offset and Scale

This ablation study investigates our method’s sensitivity to the maximum offset and maximum scale of Gaussian Primitives. For each parameter, we evaluate values from the set {0.3, 0.5, 1.0}. The results, presented in Table 2, demonstrate relatively stable performance across these configurations. Optimal performance is observed when both the maximum offset and scale are set to 0.5; consequently, these are adopted as their default values.

Table 2: Ablation study on max_offset and max_scale on KITTI dataset.

Max_Offset(m)	Max_Scale(m)	λ_1	Same Area		Cross Area	
			Mean(m) ↓	Median(m) ↓	Mean(m) ↓	Median(m) ↓
0.3	0.3	0	6.16	2.89	7.36	3.20
0.5	0.5	0	5.82	2.85	7.05	3.22
1.0	1	0	6.00	2.95	7.06	3.16
0.3	0.3	1	3.42	2.28	6.83	2.53
0.5	0.5	1	2.87	2.06	6.20	2.51
1.0	1	1	3.28	2.30	6.52	2.57

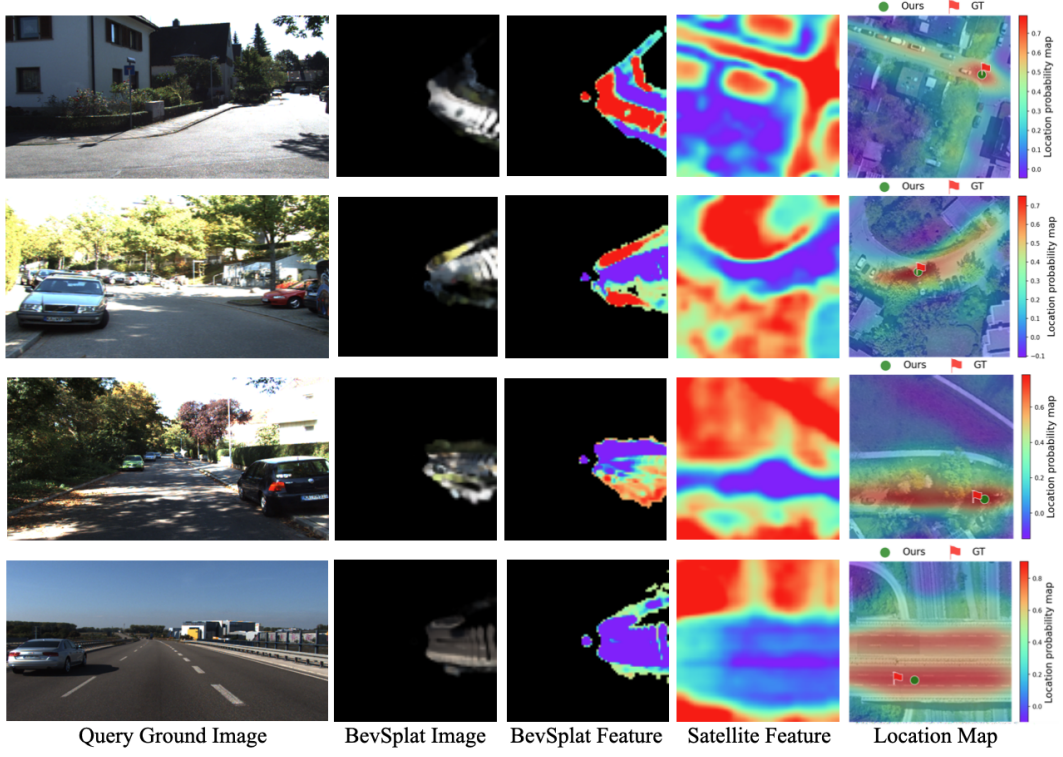


Figure 1: Qualitative results for BevSplat-based single-image localization on KITTI. Top two rows: successful examples; bottom two rows: failure examples.

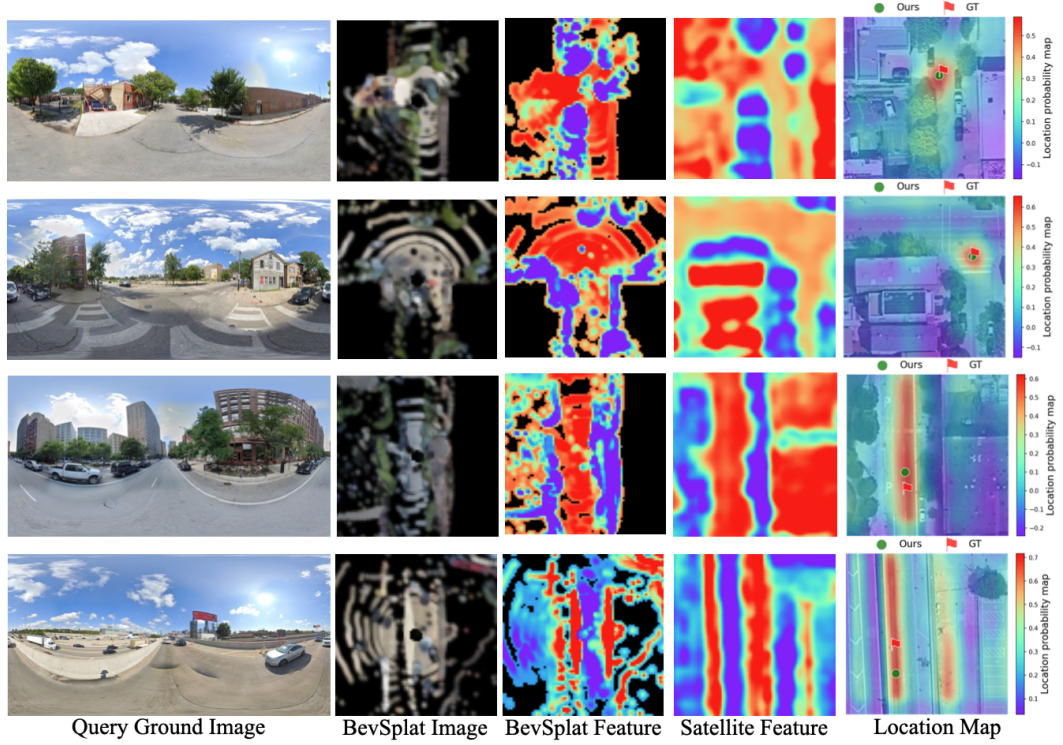


Figure 2: Qualitative results for BevSplat-based single-image localization on VIGOR. Top two rows: successful examples; bottom two rows: failure examples.

3 Qualitative Results

Robust Performance in Complex Scenarios: Our method demonstrates robust localization performance across a variety of challenging scenarios, such as road intersections, curved road sections, and areas with significant occlusions from roadside trees, as validated on the KITTI and VIGOR datasets. This proficiency is primarily attributed to our approach’s enhanced capabilities in: (1) effectively handling visual occlusions caused by buildings; (2) establishing and leveraging more accurate geometric relationships within the scene; and (3) optimally fusing features pertinent to the vertical spatial arrangement of elements, such as trees and road surfaces, between ground-level and aerial (*e.g.*, satellite) views. Consequently, our method achieves promising localization results in these complex environments, underscoring its effectiveness in tackling real-world complexities.

Limitations in Feature-Scarce Environments: Conversely, in specific scenarios such as long, straight road segments that lack distinctive visual features, our method exhibits a comparative reduction in localization accuracy. The primary reason for this limitation is that in the absence of salient visual landmarks, the deep learning network, when attempting to match the ground-level view to the satellite imagery, may assign similar matching probabilities or confidence scores to multiple plausible locations within the satellite map. This multi-modal matching outcome leads to localization ambiguity, making it difficult for the network to make a unique, high-precision positioning decision.

4 Applicability to Multi-Frame Localization Tasks

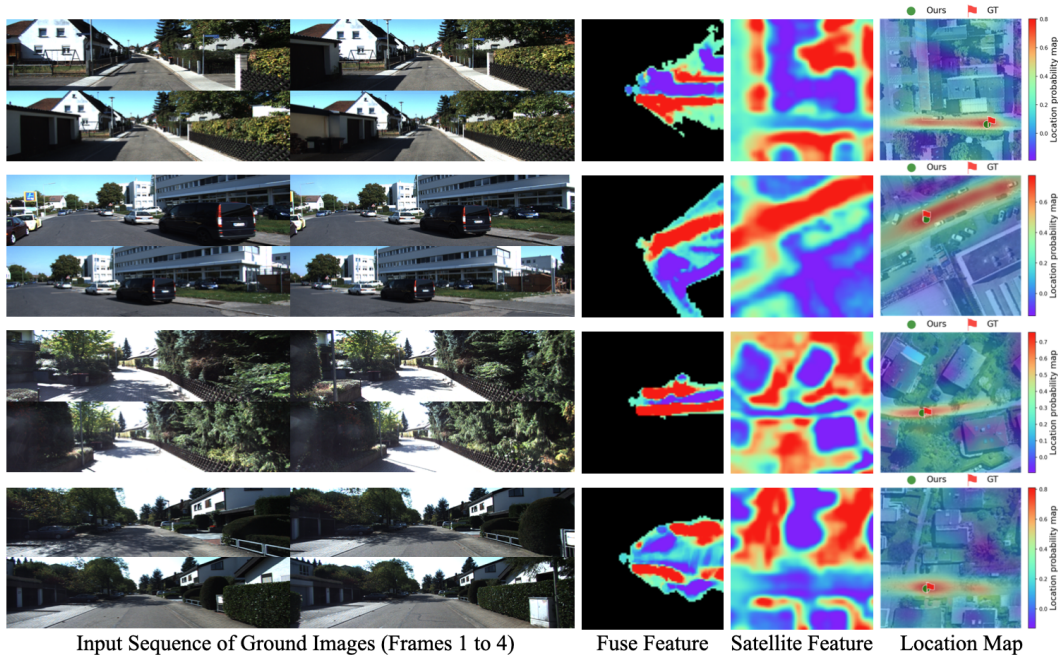


Figure 3: Visualization of multi-frame localization results on the KITTI dataset, achieved using our BevSplat-based approach. This demonstrates the aggregation of information over time and the filtering of dynamic elements.

As detailed in the main paper, our quantitative analysis of a CVLNet-based multi-frame fusion method [1] demonstrated progressively improved performance with an increasing number of frames, confirming its efficacy for temporal sequence tasks. To complement these findings, Figure 3 provides a qualitative illustration.

This visualization highlights how our fusion strategy effectively aggregates richer contextual information from multiple video frames. Notably, the approach adeptly filters dynamic objects while prioritizing the preservation of static scene elements, which are crucial for robust cross-view localization. These qualitative insights further substantiate the effectiveness and generalizability of our proposed method in handling dynamic environments and leveraging temporal information for improved localization.

5 Coordinate System

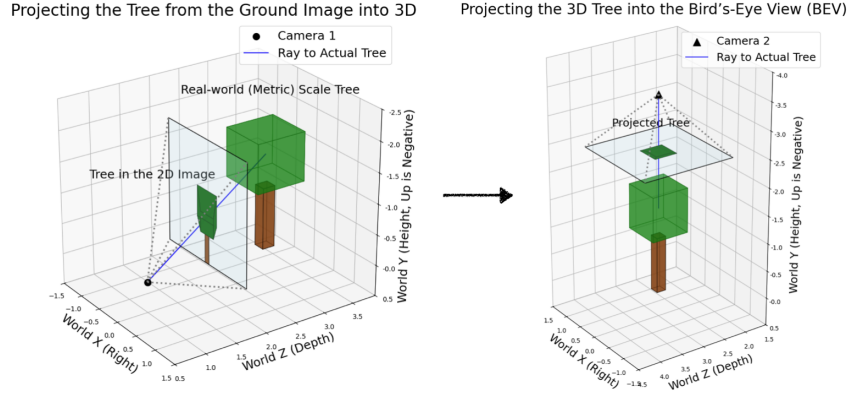


Figure 4: BevSplat Geometry Projection Overview. Our method is a two-stage geometry projection. *Left panel (Stage 1):* We reconstruct the 3D scene from ground-level images using their associated depth information, illustrated by converting a tree from a ground-level image to its 3D representation. *Right panel (Stage 2):* The reconstructed 3D scene is then projected into the Bird’s-Eye View (BEV).

Our methodology employs a world coordinate system consistent with the OpenCV convention [2], as depicted in Figure 4. This is a right-handed system where, from the camera’s viewpoint, the +X axis extends to its right, the +Y axis points downwards, and the +Z axis aligns with its forward viewing direction. Consequently, the upward direction corresponds to the -Y axis.

In the 3D reconstruction stage, a point cloud is generated from the input images. This is achieved by back-projecting pixels, using their depth information, along the initial camera’s viewing direction (defined as the +Z axis of this coordinate system).

Subsequently, for Bird’s-Eye View (BEV) projection, an aerial perspective is simulated. A virtual camera is conceptually positioned at a nadir viewpoint—looking directly downwards—above the reconstructed 3D scene. Given our coordinate system where the +Y axis points downwards, this BEV camera is located at a Y-coordinate that is numerically smaller than those of the scene’s primary content (thus representing a higher altitude). It views along the +Y direction (downwards). The BEV is then formed by orthographically projecting the 3D point cloud onto the world’s XZ-plane (which effectively serves as the ground plane) along this +Y viewing axis.

6 Limitations and Future Works

As discussed in the main paper’s conclusion, a current limitation of our BevSplat method, which renders Bird’s-Eye View (BEV) perspectives based on 3D Gaussian Splatting [3], is its computational speed compared to Inverse Perspective Mapping (IPM). For instance, on an NVIDIA RTX 4090 GPU, BevSplat requires 14 ms to generate a single BEV image. In contrast, IPM, which utilizes direct linear interpolation, can achieve this in 4ms. This performance disparity affects the overall inference speed of our model. Therefore, a significant direction for our future work is the exploration of faster and more compact Gaussian representations to address this bottleneck and enhance real-time applicability.

7 Broader Impacts

Our work, BevSplat, addresses the critical demand for robust and accessible localization systems for mobile robots, such as drones and autonomous vehicles, particularly in scenarios where high-precision GPS is either unavailable or impractical due to cost or signal dependency. By leveraging computer vision, BevSplat delivers real-time, high-precision localization using only a monocular camera, or a camera augmented with low-cost, low-precision GPS. This significantly extends localization capabilities to GPS-denied or unreliable environments, a crucial step for the widespread adoption of autonomous systems.

70 To foster further research and collaboration within the community, we are committed to open-sourcing
71 our complete codebase, training datasets, and pre-trained model weights on GitHub. This efficient
72 implementation, which operates on a single NVIDIA RTX 4090 GPU, is provided as a resource for
73 the research community. We encourage researchers to explore, build upon, and collaborate with
74 us to advance this promising research direction, ultimately contributing to safer and more versatile
75 autonomous navigation.

76 **References**

- 77 [1] Y. Shi, X. Yu, S. Wang, and H. Li, “Cvlnet: Cross-view semantic correspondence learning for video-based
78 camera localization,” in *Asian Conference on Computer Vision*. Springer, 2022, pp. 123–141.
- 79 [2] G. Bradski, “Learning opencv: Computer vision with the opencv library,” *O’REILLY google schola*, vol. 2,
80 pp. 334–352, 2008.
- 81 [3] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field
82 rendering,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.