Supplementary Material for "DisCo-FLoc: Using Dual-Level Visual-Geometric Contrasts to Disambiguate Depth-Aware Visual Floorplan Localization"

Shiyong Meng[1] , **Tao Zou**[1] , **Bolei Chen**[1*] , **Chaoxu Mu**[2] , **Jianxin Wang**[1]

[1]School of Computer Science and Engineering, Central South University
[2]School of Artificial Intelligence, Anhui University
{xiaowugui1017, 244711024, boleichen}@csu.edu.cn, cxmu@tju.edu.cn, jxwang@mail.csu.edu.cn

1 More Experimental Details

1.1 Implementation of 3DP & RSK

3DP & RSK is achieved by integrating the visual encoders pre-trained in 3DP [Chen *et al.*, 2025a] and RSK [Chen *et al.*, 2025b] into a unified Floc framework. As shown in Fig. 1 (a), the fully pre-trained encoders, F_θ in 3DP and F_ϑ in RSK, are transferred to the visual FLoc framework for fine-tuning to fit the task, which localizes by finding the pose in the floorplan that has the most similar 2D rays (similar to LIDAR scans) as the prediction. Our visual FLoc framework is implemented as a dual-branch model consisting of a 3D geometric prior branch and a RSK branch. In each branch, the image is first aligned with the gravity direction, as done in [Chen *et al.*, 2024]. Then, F_θ/F_ϑ and an attention [Vaswani, 2017] based network are used to learn the probability distribution of planar depth over a range of depth hypotheses. Pixels that become unobservable due to gravity alignment are masked in the attention, as shown in Fig. 1 (b). To adaptively leverage 3D geometric priors and RSK based on the current view, a selection network (as shown in Fig. 1 (c)) implemented as a multilayer perceptron is adopted to learn a weight $0 \leq \omega \leq 1$ from the two predictions for adaptive selection:

$$\mathbf{P}_{Fusion} = \omega \mathbf{P}_{3DP} + (1 - \omega)\mathbf{P}_{RSK}. \tag{1}$$

$\mathbf{P}_{3DP}$ and $\mathbf{P}_{RSK}$ denote the probability distributions of planar depth from the 3D geometric prior branch and the RSK branch, respectively. The expectation of $\mathbf{P}_{Fusion}$ provides the final prediction of 2D rays. ω is manually specified as 1 and 0 implying that only 3D geometric priors and RSK are used, respectively. For the training of FLoc models, we optimize an L1 loss and a cosine similarity-based shape loss:

$$\mathcal{L}_{FLoc} = ||\mathbf{d}, \mathbf{d}^*||_1 + \frac{\mathbf{d}^\top \mathbf{d}^*}{max\{||\mathbf{d}||_2 ||\mathbf{d}^*||_2, \epsilon\}}, \tag{2}$$

where $\mathbf{d}$ and $\mathbf{d}^*$ are predicted and GT 2D-ray depths, respectively. ϵ is a small constant to prevent division by zero.

1.2 Variants of SemRayLoc

Similar to the implementation of 3DP & RSK in Subsection 1.1, **SemRayLoc**$_r$ **+ 3DP** is implemented by replacing the depth and semantic ray encoders in Fig. 2 with the fully pre-trained visual encoder F_θ from 3DP. Similarly, **SemRayLoc**$_r$ **+ RSK** is implemented by replacing the depth and semantic ray encoders in Fig. 2 with the fully pre-trained visual encoder F_ϑ from RSK. **SemRayLoc**$_r$ **+ 3DP & RSK** is implemented by replacing the depth and semantic ray encoders with F_θ and F_ϑ, respectively.

Table 1: Comparative studies of long-sequence trajectory tracking methods on the Gibson(t) dataset.

Method (Venue)	**Gibson(t) R@**			
	0.2 m↑	**1 m↑**	**RMSE(S)↓**	**RMSE(A)↓**
LASER(CVPR 2022)	-	59.5	0.39	1.96
F^{3}Loc(CVPR 2024)	35.1	89.2	0.18	0.88
F^{3}Loc fusion(CVPR 2024)	62.2	94.6	0.12	0.51
3DP(ACM MM 2025)	70.3	97.3	0.12	0.34
RSK(AAAI 2026)	59.5	94.6	0.13	0.51
3DP & RSK	64.9	94.6	0.12	0.51
Ours (DisCo-FLoc)	73.0	94.6	0.11	0.49
Oracle	-	100.0	0.07	0.07

2 Additional Evaluation for Long-Sequence Trajectory Tracking

In this section, we compare our method with existing SOTA methods on the long-sequence trajectory tracking task using the Gibson(t) dataset collected by F^{3}Loc [Chen *et al.*, 2024]. Gibson(t) consists of 118 pieces of long-sequence views, each of which contains 280 $\sim$ 5152 image frames. The **R**oot-**M**ean-**S**quare **E**rror (RMSE) (over the last 10 frames) is employed to measure the accuracy of sequential trajectory tracking when localization is successful (RMSE(S)) and in all cases (RMSE(A)). Technically, we combine the histogram filter proposed by F^{3}Loc with our method and perform visual FLoc using 100 historical frames.

As shown in Tab. 1, our method improves the Recall metric by 2.7% compared to 3DP at the localization accuracy of 0.2 m. The reduction in the RMSE(S) metric reflects the robustness of our method in sequential trajectory tracking. Additionally, our method achieves competitive results in terms of the localization accuracy of 1 m and RMSE(A). It is worth noting that F^{3}Loc fusion performs visual FLoc by adaptively utilizing single-frame and multi-frame images, yet it is only competitive on the RMSE(S) metric. Oracle achieved 100% Recall at 1 m accuracy by combining GT rays with a his-

[*]Corresponding Author.

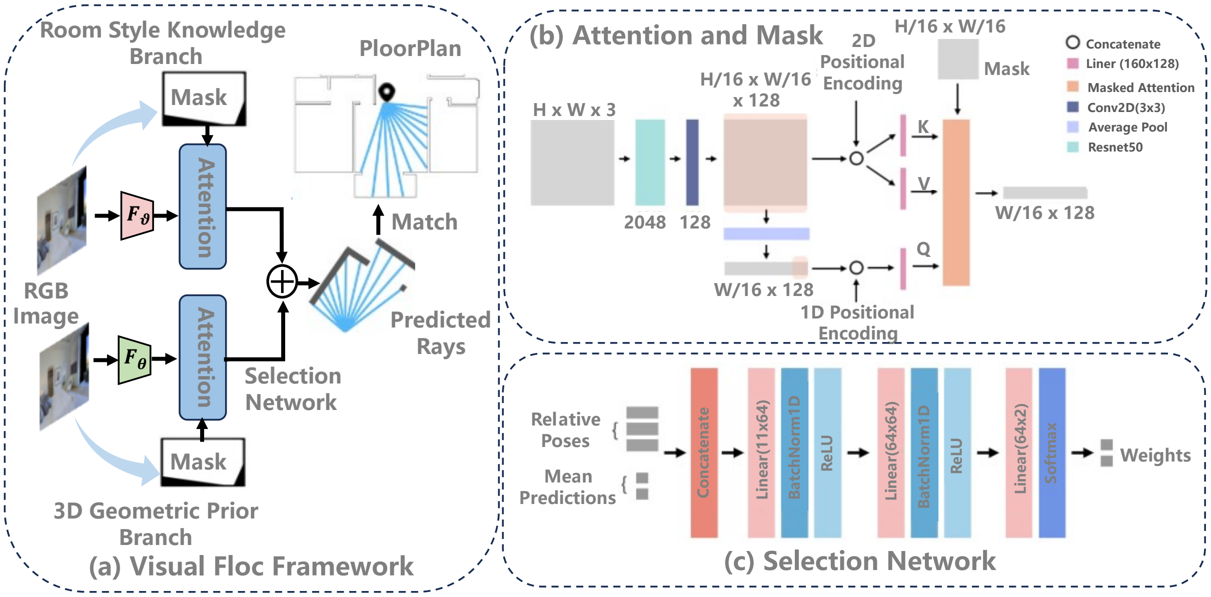

Figure 1: Illustrations of 3DP & RSK. (a) The pre-trained visual encoders, F_θ in 3DP and F_ϑ in RSK, are transferred to the visual FLoc framework for fine-tuning to further fit this task. The visual FLoc framework is a dual-branch model consisting of a 3D geometric prior branch and a RSK branch. (b) and (c) detail the masked attention mechanism and the selection network, respectively.

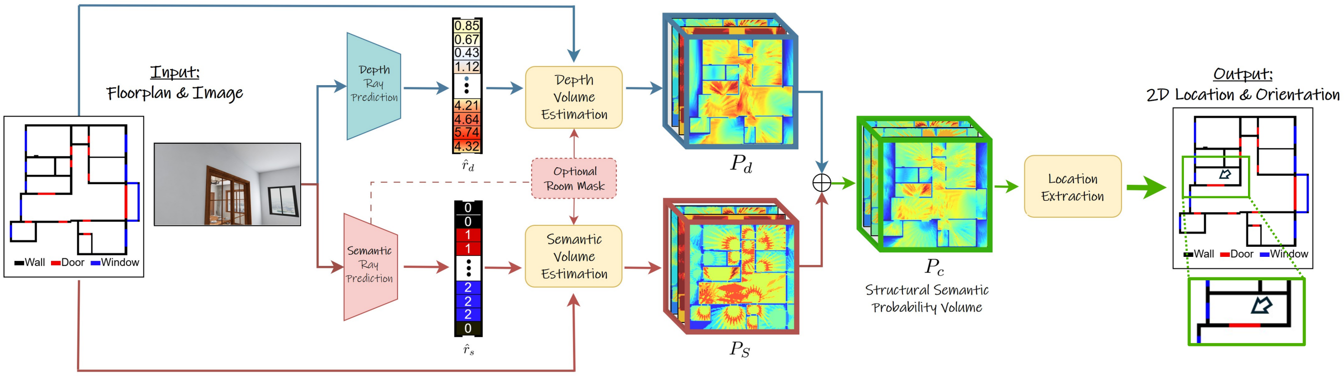

Figure 2: The pipeline of SemRayLoc from the paper [Grader and Averbuch-Elor, 2025].

togram filter, demonstrating the potential of visual FLoc.

3 More Visualizations

Fig. 3 and Fig. 4 illustrate the qualitative comparisons between our method and SemRayLoc [Grader and Averbuch-Elor, 2025]. Influenced by the repetitive geometric structures in the floorplan, our RRP also fails when SemRayLoc encounters errors. However, our visual-geometric contrasts effectively address localization ambiguities and yield high-precision final FLocs.

References

[Chen *et al.*, 2024] Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F3loc: Fusion and filtering for floorplan localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18029–18038, 2024.

[Chen *et al.*, 2025a] Bolei Chen, Jiaxu Kang, Haonan Yang, Ping Zhong, and Jianxin Wang. Perspective from a higher dimension: Can 3d geometric priors help visual floorplan localization? In *Proceedings of the 33nd ACM International Conference on Multimedia*, 2025.

[Chen *et al.*, 2025b] Bolei Chen, Shengsheng Yan, Yongzheng Cui, Jiaxu Kang, Ping Zhong, and Jianxin Wang. Perspective from a broader context: Can room style knowledge help visual floorplan localization? *arXiv preprint arXiv:2508.01216*, 2025.

[Grader and Averbuch-Elor, 2025] Yuval Grader and Hadar Averbuch-Elor. Supercharging floorplan localization with semantic rays. *arXiv preprint arXiv:2507.09291*, 2025.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

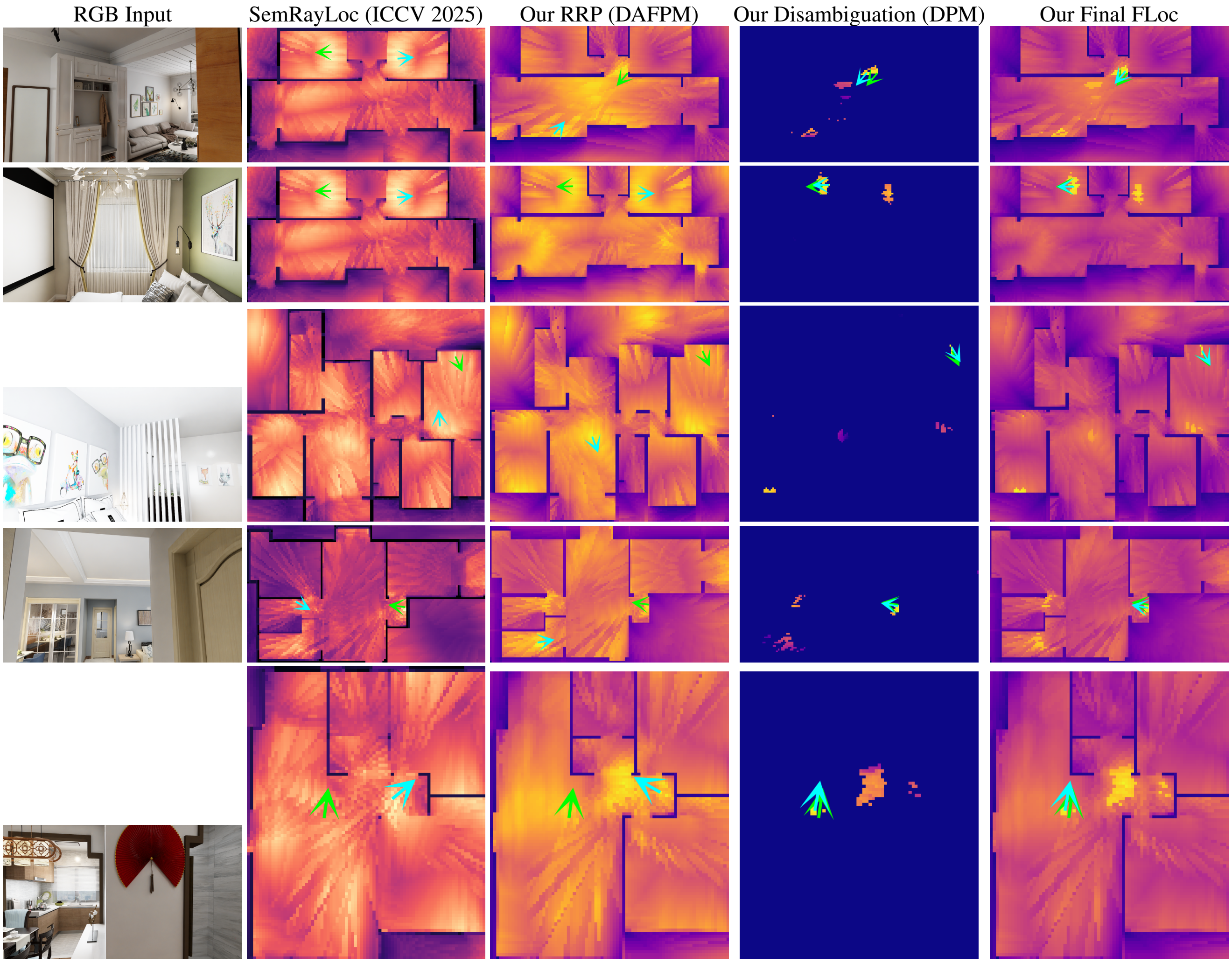

Figure 3: Qualitative comparison on 9 scenes (Part 1/2).

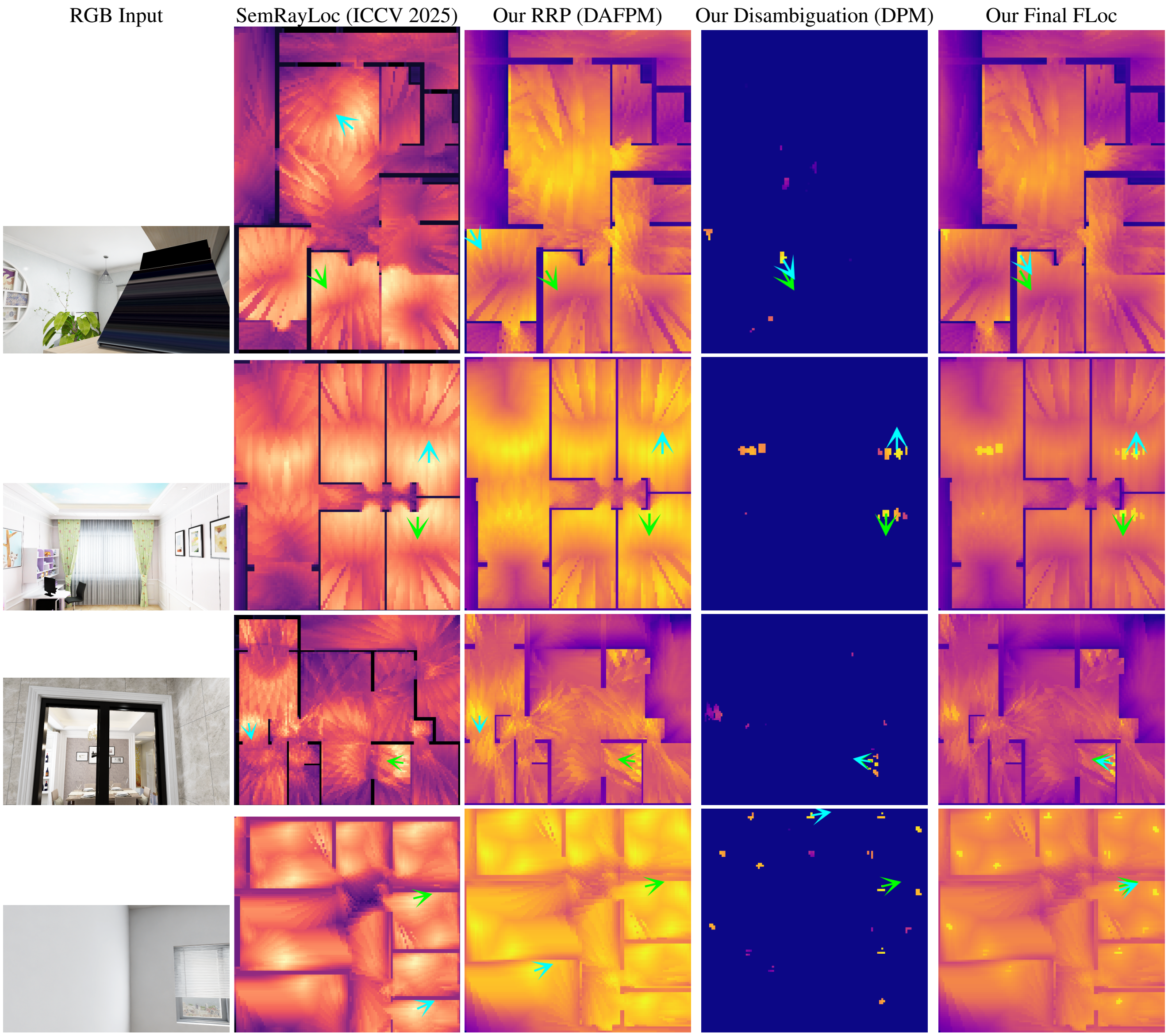

Figure 4: Qualitative comparison on 9 scenes (Part 2/2, Continued).