Supplementary Material for
PALMS+: Modular Image-Based Floor Plan Localization Leveraging Depth Foundation Models

Yunqian Cheng *Benjamin Princen* *Roberto Manduchi*
University of California, Santa Cruz, United States
{ychen827, bprincen, manduchi}@ucsc.edu

A. Technical Details

A.1. Layout Matching

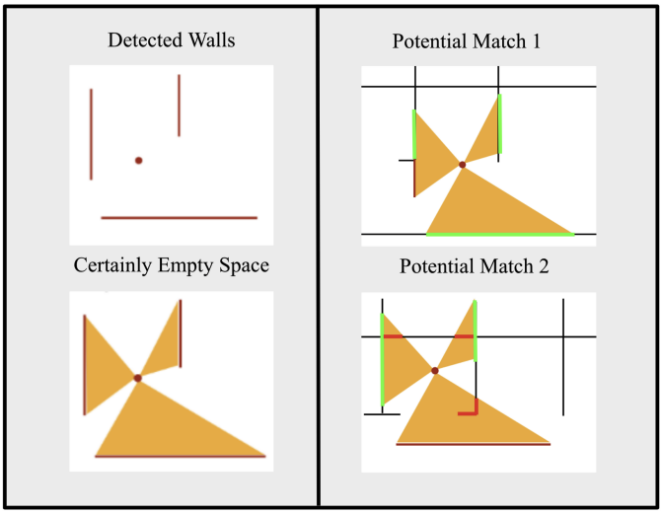


Figure 6. Illustration of the Certainly Empty Space (CES) from [5]. It shows a potentially good match (top right) and a bad match (bottom right) because of the floor plan segments' intersection with the CES, which breaks the visibility constraint.

Given observed wall segments, we employ a convolution-based approach to estimate the likelihood that a particular location corresponds to the current position. The convolution kernel is constructed by rasterizing the observed wall segments. To account for uncertainty in the orientation, the vector map of wall segments is rotated O times, yielding O distinct kernels.

The orientations of wall segments are extracted directly from their vector form and normalized to the interval $[-\pi/2, \pi/2]$. The same process is done on the vector floor plan, then we find the orientation candidates using the method illustrated in Figure 7. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_5\}$ denote the top five estimated orientation candidates. Because wall segment orientations are ambiguous up to π (i.e., a vector direction cannot distinguish between θ and $\theta + \pi$), we augment this set by adding π to each element of Θ, producing

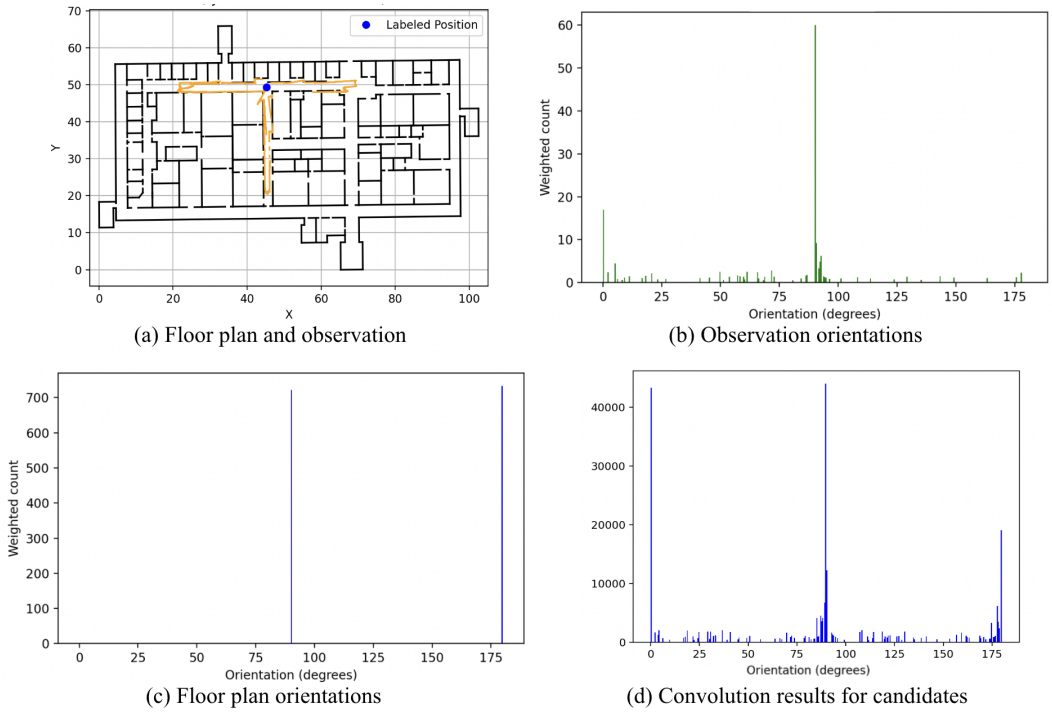
$$\Theta' = \{\theta_1, \ldots, \theta_5, \theta_1 + \pi, \ldots, \theta_5 + \pi\}.$$

The convolution kernels are then constructed by rotating the vector map according to each $\theta \in \Theta'$, resulting in $O = 10$ distinct kernels. The corresponding convolution operations yield spatial distributions of likelihood values across the floor plan.

However, this baseline approach does not inherently enforce visual occlusion constraints. If a wall is observed, no additional walls should be detected between the observer and that wall. This principle, termed *Certainly Empty Space* (CES), leverages the geometry of observed walls to define regions that are guaranteed to remain unoccupied. Concretely, the triangular region subtended by an observed wall segment and the observer's position defines an exclusion zone within which no additional wall segments should appear. Each exclusion zone is defined by three points: the observer's location and the two endpoints of the wall segment. These exclusion zones, together with the vector map, are rotated according to the orientations in Θ' and rasterized to obtain the final set of convolution kernels. CES is illustrated in Figure 6.

A.2. Runtime Analysis

We measured the runtime of each component in our pipeline, from point cloud reconstruction to the final posterior computation, on a device equipped with an Apple M2 chip. The point cloud reconstruction and scale alignment steps took 20.447



(a) Floor plan and observation (b) Observation orientations

(c) Floor plan orientations (d) Convolution results for candidates

Figure 7. Illustration of the candidate orientation extraction process. For each set of line segments, we extract the orientations and weight each with the segment length.

seconds, while projecting the point cloud to line segments and determining their candidate orientations each required 0.005 seconds. The final posterior estimation took 3.773 seconds.

Overall, the current implementation is not yet suitable for real-time deployment. However, there is substantial room for optimization. The runtime could be reduced through code-level optimization, lowering image or point cloud resolution, or leveraging GPU acceleration. We exclude the runtime of the depth foundation model, as its performance metrics are reported in the original paper.

B. Additional Experimental Results

We show additional experimental results from PALMS [5] and F^{3}Loc [4] as well as our own results. These results, as well as results from Tables 1 and 2 of the main paper, are from the same experiments.

We note that for the custom dataset, the localization accuracy is the same for @1m and @1m30°. This is expected because our dataset emphasizes transitional spaces—hallways, intersections, and lobbies—that provide richer geometric and directional cues. As a result, when the position estimate is accurate (within 1 m), the inferred orientation is almost always consistent (within 30°) as well.

C. Additional Qualitative Examples

D. Dataset Specs

D.1. Privacy

To protect privacy, all images are preprocessed to remove sensitive content. Individuals appearing in photographs were manually blurred out to prevent inadvertent identification, and computer screens were similarly blurred to avoid inadvertent exposure of private or confidential information.

D.2. Distribution of Observation Points

See Figure 11.

Building	Method	Localization Accuracy (%)					
		0.1m	0.5m	1m	1m 30°	2m	5m
1	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	15.8	26.3	26.3	26.3	26.3	26.3
	PALMS+	0.0	47.4	52.6	52.6	57.9	57.9
	PALMS+*	0.0	52.6	57.9	57.9	68.4	68.4
2	F^3Loc	0.0	0.0	0.0	0.0	0.0	4.2
	PALMS	0.0	0.0	11.1	11.1	0.0	0.0
	PALMS+	0.0	8.3	8.3	8.3	8.3	8.3
	PALMS+*	4.2	8.3	12.5	12.5	12.5	16.7
3	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS+	0.0	11.1	33.3	33.3	44.4	50.0
	PALMS+*	0.0	33.3	50.0	50.0	50.0	55.6
4	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	5.6	5.6	5.6	5.6	5.6	11.1
	PALMS+	0.0	16.7	33.3	33.3	33.3	33.3
	PALMS+*	5.6	22.2	38.9	38.9	38.9	38.9

Table 5. Localization accuracy on the custom dataset under the full-view observation setting. This table shows per-building metrics for all methods. PALMS+* is PALMS+ with masked depths.

Building	Method	Localization Accuracy (%)					
		0.1m	0.5m	1m	1m 30°	2m	5m
1	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	5.3	10.5	10.5	10.5	10.5	15.8
	PALMS+	0.0	26.3	31.6	31.6	31.6	31.6
	PALMS+*	0.0	26.3	36.8	36.8	42.1	47.4
2	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS+	0.0	4.2	12.5	12.5	12.5	12.5
	PALMS+*	0.0	8.3	12.5	12.5	12.5	12.5
3	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS+	0.0	5.6	11.1	11.1	16.7	16.7
	PALMS+*	0.0	0.0	11.1	11.1	11.1	11.1
4	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	5.6	11.1	11.1	11.1	11.1	11.1
	PALMS+	5.6	11.1	22.2	22.2	22.2	22.2
	PALMS+*	0.0	16.7	27.8	27.8	33.3	33.3

Table 6. Localization accuracy on the custom dataset under the partial-view observation setting. This table shows per-building metrics for all methods. PALMS+* is PALMS+ with masked depths.

Building	**Method**	Localization Accuracy (%)					
		0.1m	0.5m	1m	1m 30°	2m	5m
1	F^3Loc	0.0	0.0	0.0	0.0	0.0	5.3
	PALMS	0.0	5.3	5.3	5.3	5.3	15.8
	PALMS+	5.3	5.3	15.8	15.8	15.8	15.8
	PALMS+*	0.0	5.3	21.1	21.1	21.1	26.3
2	F^3Loc	0.0	0.0	0.0	0.0	0.0	4.2
	PALMS	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS+	0.0	4.2	4.2	4.2	4.2	4.2
	PALMS+*	0.0	8.3	8.3	8.3	8.3	8.3
3	F^3Loc	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS+	0.0	5.6	5.6	5.6	5.6	11.1
	PALMS+*	0.0	5.6	5.6	5.6	5.6	11.1
4	F^3Loc	0.0	0.0	0.0	0.0	5.6	5.6
	PALMS	0.0	0.0	0.0	0.0	0.0	0.0
	PALMS+	0.0	11.1	22.2	22.2	27.8	27.8
	PALMS+*	0.0	16.7	22.2	22.2	22.2	27.8

Table 7. Localization accuracy on the custom dataset under the single-view observation setting. This table shows per-building metrics. PALMS+* is PALMS+ with masked depths.

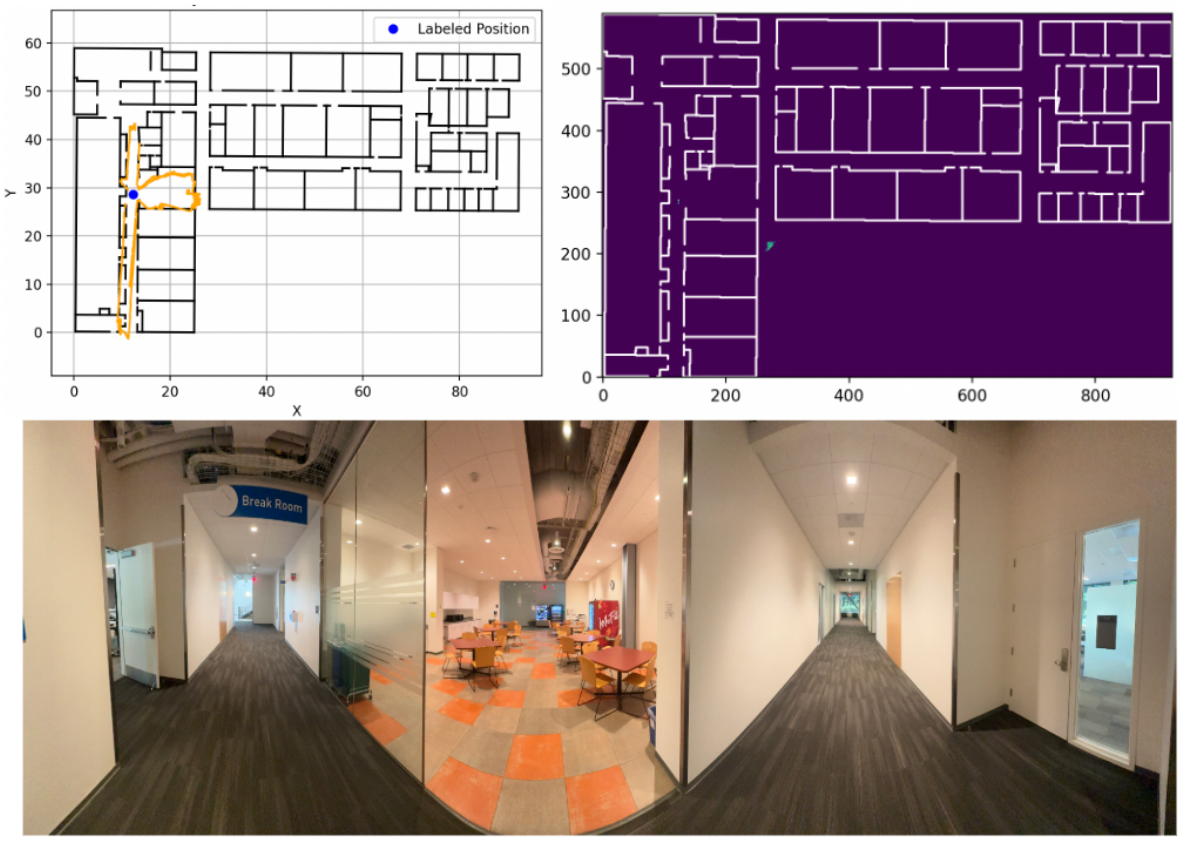


Figure 8. Distinctive structure is well captured by the system, resulting in relatively optimal performance where only a couple of peaks are present. A strong, concentrated peak at the ground-truth location can be seen in this example.

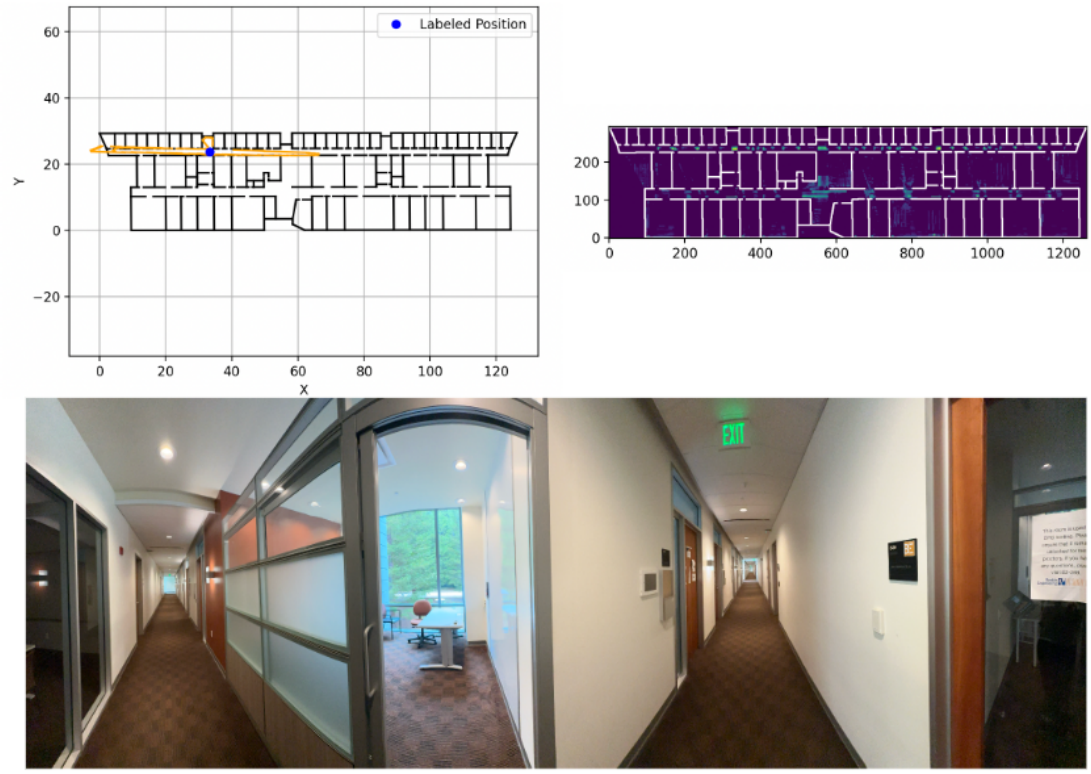


Figure 9. Long hallways are common, which results in less optimal performance compared to Figure 8. Multiple peaks are present, reflecting higher geometric ambiguity.

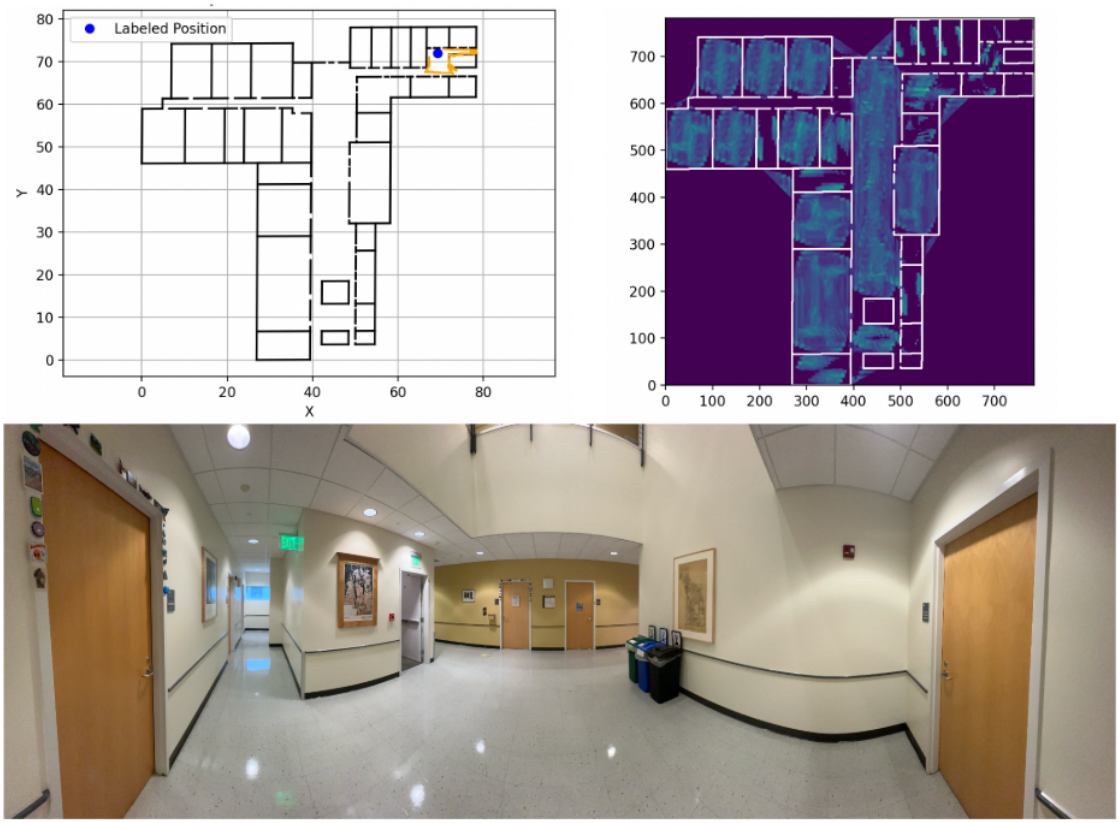
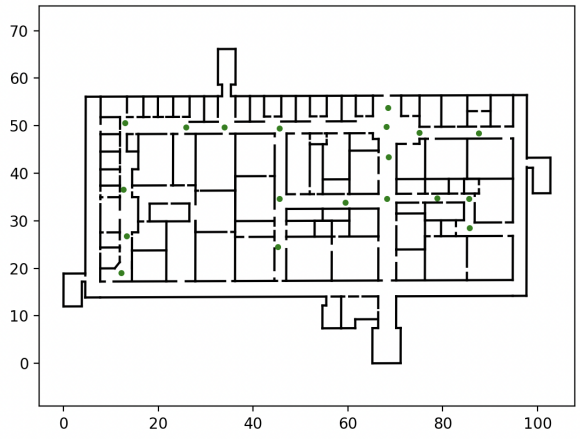
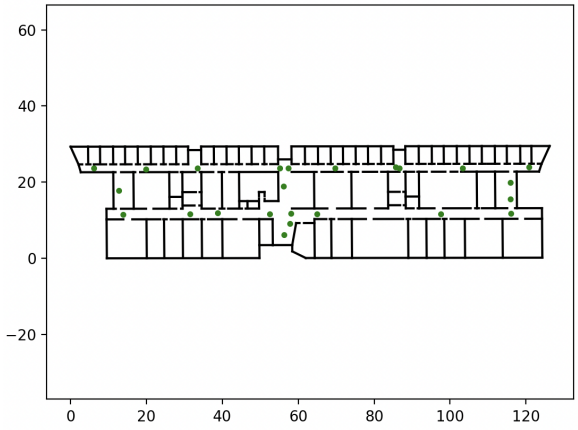


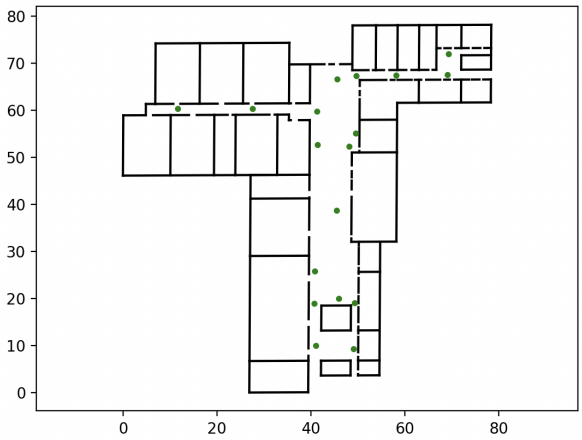
Figure 10. Observation exhibits a simple box-like structure, which results in low performance and higher ambiguity, as shown by the heatmap.



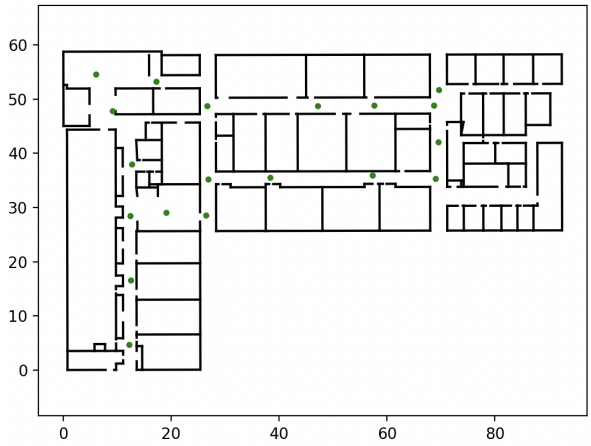
(a) Building 1



(b) Building 2



(c) Building 3



(d) Building 4

Figure 11. The figures above show all the points at which data were recorded in each of the buildings. We intentionally included a mixture of locations with different levels of ambiguity. Building 2 is the most difficult to localize among all the buildings because it consists mostly of long, homogeneous hallways.