Render-Rank-Refine: Accurate 6D Indoor Localization via Circular Rendering

Haya Monawwar and Guoliang Fan *

Posted Date: 14 November 2025

doi: 10.20944/preprints202511.1027.v1

Keywords: indoor localization; layout ambiguity; pose estimation; rotation-invariant descriptors; semantic models

 doi:10.20944/preprints202511.1027.v1

Article

Render-Rank-Refine: Accurate 6D Indoor Localization via Circular Rendering

Haya Monawwar and Guoliang Fan *

School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, Oklahoma, USA
* Correspondence: guoliang.fan@okstate.edu

Abstract

Accurate six-degree-of-freedom (6-DoF) camera pose estimation is essential for augmented reality, robotics navigation, and indoor mapping. Existing pipelines often depend on detailed floorplans, strict Manhattan-world priors, and dense structural annotations, which may lead to failures in ambiguous, overlapping-room layouts (ambiguous? not overlapping). We present *Render-Rank-Refine*, a two-stage framework operating on coarse semantic meshes without requiring textured models or per-scene fine-tuning. First, panoramas rendered from the mesh enable global retrieval of coarse pose hypotheses. Then, perspective views from the top-k candidates are compared to the query via rotation-invariant circular descriptors, which reranks the matches before final translation and rotation refinement. In general, our method reduces the translation and rotation error by an average of 40% and 29%, respectively, compared to the baseline while achieving more than 90% improvement in cases with severe layout ambiguity. It sustains 25–27 queries per second (QPS), which is about 12 times faster than the existing state-of-the-art, without sacrificing accuracy. These results demonstrate robust, near-real-time indoor localization that overcomes structural ambiguities and heavy geometric assumptions.

Keywords: indoor localization; layout ambiguity; pose estimation; rotation-invariant descriptors; semantic models

1. Introduction

Indoor camera localization, which is estimating a camera's 6-degree-of-freedom (6-DoF) pose in complex indoor spaces, is a core problem in computer vision with applications in navigation [1,2], augmented reality [3], and assistive technologies for people with disabilities [4]. Classical Structure-from-Motion and geometric methods [5,6] achieve high accuracy under controlled conditions but assume Manhattan world layouts or dense three-dimensional reconstructions. On the other hand, learning-based approaches such as PoseNet [7] offer greater flexibility but struggle in visually ambiguous scenarios. Moreover, traditional methods rely on dense scene-specific data including curated images, depth maps, or three-dimensional point clouds, limiting scalability. Lastly, although synthetic augmentation and simplified three-dimensional representations such as Structured3D [8] improve applicability, most methods focus on two-dimensional localization, which still leaves critical degrees of freedom unmodeled for 6-DoF poses.

Recent methods such as SPVLoc [9] and LASER [10] have advanced indoor pose estimation. SPVLoc does so by directly linking perspective queries to rendered semantic panoramas and achieves robust localization in unseen indoor layouts while LASER builds on this idea with a Monte Carlo inference pipeline and a geometrically organized latent space, rendering circular descriptors from floor-plan codebooks to deliver efficient and precise indoor localization. However, the assumption of largely unobstructed views, and correlation-based refinement is sensitive to the top-ranked panorama, so a single retrieval error can derail the estimate. Furthermore, heavy reliance on detailed and accurate floorplan annotations (such as window placements) and the 2D abstraction can limit performance under incomplete geometry or mismatched fields of view. While semantic retrieval methods provide

 doi:10.20944/preprints202511.1027.v1

scene-level robustness, they collapse under ambiguity. On the other hand, geometry-driven methods provide invariances to viewpoint, orientation, and appearance changes, but may still struggle without precise window-door-opening (WDO) annotations on the room layout. Furthermore, accurate floorplan registration is rarely feasible in practice, especially when on-the-fly pose estimation is needed in safety-critical settings such as caring for people with disabilities [4]. Hence, reliable camera localization is generally more possible in environments, such as Figure 1(a,b), due to clear textures and distinctive room geometry. However, when multiple locations exhibit near-identical visual cues, as in Figure 1(c,d), localization becomes substantially more difficult, transforming layout ambiguity from an occasional nuisance into a persistent obstacle for robust 6-DoF estimation. Scenes that capture two or more rooms within a single view exemplify these overlapping layouts, which we term *ambiguous* or *overlapping indoor environments*.

Figure 1. ZInD [11] examples of visually clear and ambiguous indoor scenes. (**a**) Open tiled room with clear wall–floor boundaries and window geometry, providing strong geometric cues for reliable camera localization. (**b**) Bedroom with distinctive ceiling fan, window blinds, and wood floor texture, offering unique orientation features that simplify pose estimation. (**c**) Upper-level loft visually resembling other corridors, causing ambiguity in camera pose estimation. (**d**) Small pantry adjacent to a kitchen with textures and lighting similar to other spaces, challenging localization.

In this paper, we introduce **Render-Rank-Refine**, a 6-DoF camera localization framework that requires no scene-specific training or exhaustive annotations. It handles varying camera FoVs and mitigates errors from incorrect top-ranked candidates. The method integrates semantic and geometric representations to achieve superior accuracy in ambiguous indoor environments (see Figure 1). The proposed pipeline is illustrated in Figure 2. First, a perspective query and a 3D semantic model generate top-K candidate poses via viewport matching on rendered panoramas (from (a) to (b)). Then,

 doi:10.20944/preprints202511.1027.v1

each candidate is re-rendered, encoded into 360° circular descriptors using a ResNet50 backbone (step (c)), and ranked by cosine similarity (step (d)), yielding a refined 6-DoF pose.

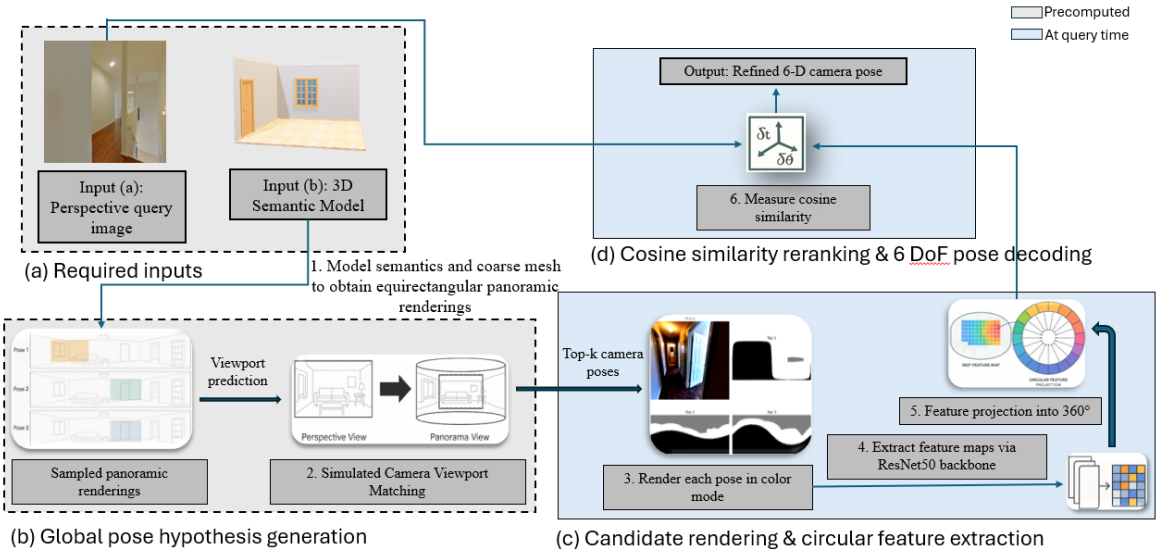


Figure 2. Overview of Render-Rank-Refine. Pre-rendered panoramas generate top-K pose hypotheses via viewport matching. Each hypothesis is re-rendered, encoded into 360° circular descriptors, and ranked by cosine similarity to produce a refined 6-DoF pose.

Our contributions are fourfold: (*i*) formalization of layout ambiguity and its connection to pose failures, enabling a principled analysis of when and why camera localization may break down, (*ii*) Bayesian refinement using mesh-segmentation retrieval as prior and rotation-invariant circular descriptors as likelihood, (*iii*) group-theoretic analysis showing marginalization over $SO(2) \subset SE(3)$ removes in-plane rotation while preserving global layout cues, and (*iv*) comprehensive evaluation on the Zillow Indoor Dataset (ZInD) [11] with ambiguity-stratified analysis. To the best of our knowledge, this is the first work to formalize and address layout ambiguity in indoor pose estimation via principled Bayesian integration of complementary representations, and achieve a highly substantial pose estimation improvement in cases with high layout ambiguity via a computationally light-weighted pipeline. Our pipeline estimates an accurate 6-DoF camera pose with minimal query-time overhead of approximately 3.5 milliseconds and maintains high performance in visually ambiguous environments.

2. Related Work

Robust 6-DoF indoor localization draws on four complementary directions: global retrieval, local features, view synthesis, and hybrid geometry-learning. Retrieval prunes candidates, local features capture spatial detail, rendering aligns real and synthetic views, and hybrid methods fuse learning with geometry. Studying these approaches exposes limitations in accurate camera pose estimation due to layout ambiguity, viewpoint variation, and retrieval errors, directly motivating our framework.

2.1. Image Retrieval-Based Localization

Global retrieval pipelines compress each image into a compact, learnable descriptor and leverage fast nearest-neighbor searches to shortlist pose candidates before any detailed matching. One such work is DenseVLAD [12], which aggregates hand-crafted local features into a single robust vector, boosting place-recognition recall across varied viewpoints. Patch-NetVLAD [13] builds on this by fusing multi-scale CNN features into one unified VLAD descriptor, markedly improving indoor retrieval accuracy. Lastly, Hierarchical Localization [14] divides the process into a coarse retrieval stage followed by fine-grained local-feature verification, reducing search complexity while preserving precision.

 doi:10.20944/preprints202511.1027.v1

Despite their efficiency in pruning candidates, these methods lack the spatial granularity needed to recover precise 6-DoF poses in visually ambiguous or overlapping room scenarios, where similar global descriptors may correspond to distinct physical locations.

2.2. Learned Local Feature Matching

Classical local descriptors such as SIFT [15], SURF [16], ORB [17], and BRISK [18], pioneered robust keypoint matching under challenging illumination and viewpoint variations. More recently, end-to-end frameworks like ASLFeat [19] have shown that jointly learning keypoint detection and description can yield superior repeatability and distinctiveness. However, accurate 6-DoF pose estimation remains dependent on geometric optimization, often using RANSAC for outlier rejection [20] and efficient PnP solvers [21]. These solvers require sufficient inlier correspondences, which can be difficult to obtain in low-texture or repetitive layout environments, especially in the presence of layout ambiguity.

2.3. Rendering and View Synthesis Methods

Synthesizing candidate views from proxy geometry helps bridge the domain gap between real RGB queries and CAD style models. SPVLoc renders semantic panoramas from coarse meshes for learned embedding matching [9], while LASER learns a latent rendering space from accurate floorplans [10]. LaLaLoc [22] and LaLaLoc++ [23] extend this with neural floorplan understanding and implicit 3D hallucination, reducing annotation needs and increasing throughput, though without explicitly modeling ambiguity. Scene reconstruction pipelines such as SLAM++ [24], LSD-SLAM [25], ORB-SLAM/ORB-SLAM2 [26,27], ElasticFusion [28], and SemanticFusion [29] produce dense, semantically annotated maps for relocalization. More recent methods include Neural Radiance Fields (NeRF) for photorealistic novel view synthesis from sparse captures [30] and ScanComplete for volumetric completion of partial scans [31]. These approaches, while robust, typically depend on extensive scanning or precise annotations, limiting deployment in unmodeled or dynamic environments.

2.4. Hybrid Geometry-Learning Approaches

Differentiable pipelines integrate learned features with geometric modules for end-to-end pose estimation. PoseNet [7] pioneered single-image 6-DoF relocalization with a CNN, MapNet [32] added geometry-aware consistency, VidLoc [33] employed LSTMs for temporal smoothing, CNN-SLAM [34] incorporated learned monocular depth, DSAC [35] made RANSAC differentiable, and SVO [36] achieved fast sparse alignment.

Existing methods fall short due to (*i*) lack of explicit modeling of layout ambiguity, (*ii*) sensitivity to in-plane rotation, and (*iii*) refinement that assumes correctness of the top retrieval. Our Render-Rank-Refine framework addresses these gaps by uniting high-recall semantic retrieval with rotation-invariant circular descriptors where retrieval scores serve as priors and circular similarity as a likelihood, enabling a Bayesian update that systematically improves localization under ambiguity.

3. Methodology

In this section, we present our framework for robust, efficient pose estimation in ambiguous indoor layouts. The pipeline consists of four main stages, two executed just once offline and two executed online in real time. The algorithm is illustrated in Figure 3. All modules are parallelized using multi-threaded computation.

 doi:10.20944/preprints202511.1027.v1

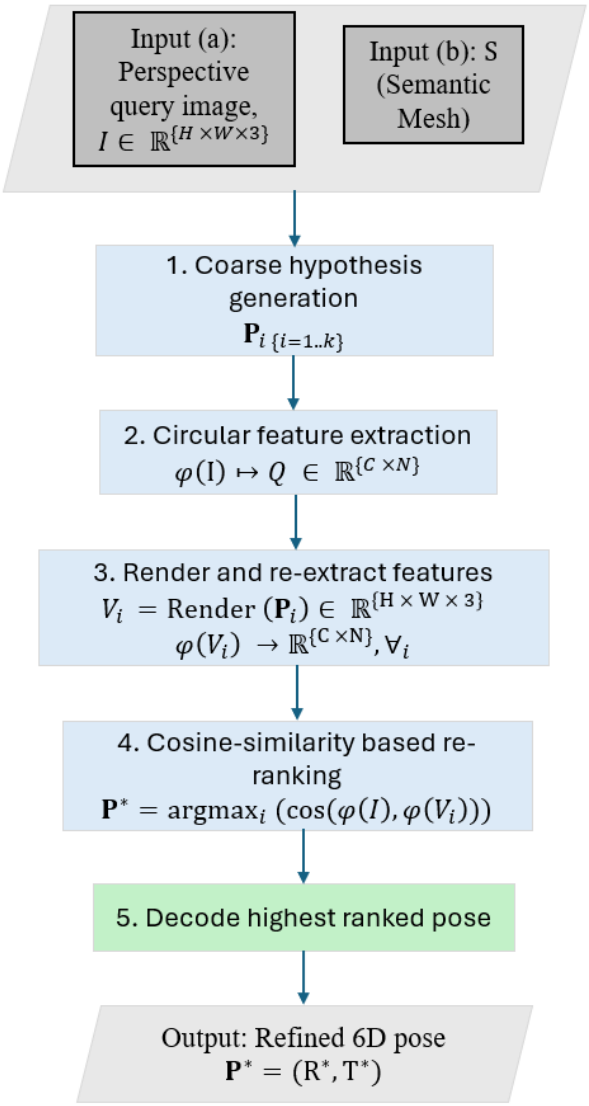


Figure 3. Overview of the proposed circular re-ranking pipeline, corresponding to the modular stages (a)-(d) illustrated in Figure 2. 1–2 (Pre-computed, modules (a)-(b)): A viewport database is generated from the semantic mesh (**Input (b)**) and candidate poses are sampled, while circular features are extracted from the perspective query (**Input (a)**). Coarse viewport matching alone may struggle in overlapping or symmetric layouts, where feature disambiguation is difficult. **3–4 (On-the-fly re-ranking, module (c)):** Each top-k candidate is rendered, circular descriptors are re-extracted, and cosine similarity is measured to refine pose ranking. **5 (Decode, module (d)):** The highest-ranked candidate is decoded into the final refined 6-DoF camera pose $(\mathbf{R}^*, \mathbf{T}^*)$.

3.1. Problem Definition

Let I_q denote a query RGB image captured at an unknown camera pose $\mathbf{P} \in SE(3)$, where $SE(3)$ represents the group of all 3D rigid-body transformations (three translational and three rotational degrees of freedom). Given a known semantic 3D model of the environment $\mathcal{S}$, the goal of indoor localization is to estimate the true pose $\mathbf{P}$ as accurately and efficiently as possible, while remaining robust to *layout ambiguity*, scenarios where different physical locations produce highly similar visual observations.

Formally, this can be expressed as a maximum a posteriori (MAP) estimation problem:

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P} \in SE(3)} p(\mathbf{P} \mid I_q, \mathcal{S}), \tag{1}$$

where $p(\mathbf{P} \mid I_q, \mathcal{S})$ is the posterior probability of a candidate pose given the query image and the known scene. This formulation emphasizes that an ideal estimator should not only retrieve the most probable

 doi:10.20944/preprints202511.1027.v1

pose but should also do so in a way that accounts for ambiguities, is computationally tractable for large-scale environments, and generalizes to diverse indoor layouts.

3.2. Quantifying Layout Ambiguity

Indoor environments often contain regions that are visually and structurally similar, such as parallel corridors, identical utility rooms, or mirrored floorplans, making them challenging to disambiguate using image descriptors alone. To reason about this analytically, we define a pose ambiguity score for a given viewpoint v as:

$$A(v) = \frac{\langle p(v), p(u) \rangle}{\|p(v)\| \cdot \|p(u)\|}, \tag{2}$$

where $f(\cdot)$ denotes the descriptor representation of a view. This score measures the highest normalized similarity between v and any other distinct viewpoint u in the environment.

A high $A(v)$ indicates that the view shares strong visual or structural resemblance with at least one other location, which increases the likelihood of retrieval errors and pose estimation failures. Empirical testing shows that failure rates increase significantly with the similarity ratio $A(v)$, confirming its value as a predictor of localization difficulty.

3.3. Theoretical Motivation: A Bayesian View

Our approach treats pose estimation not as a single-step matching problem, but as a sequential process of hypothesis generation and hypothesis verification. This naturally lends itself to a two-stage Bayesian inference formulation, in which each stage contributes complementary information about the unknown camera pose.

Formally, we express the posterior distribution over the 6-DoF pose $\mathbf{P} \in SE(3)$, given a query image I_q and a semantic 3D model $\mathcal{S}$, as:

$$p(\mathbf{P} \mid I_q, \mathcal{S}) \propto p(I_q \mid \mathbf{P}, \mathcal{S}) \cdot p(\mathbf{P} \mid \mathcal{S}). \tag{3}$$

Here, $p(\mathbf{P} \mid \mathcal{S})$ reflects our prior belief about plausible poses before observing I_q, while $p(I_q \mid \mathbf{P}, \mathcal{S})$ measures how well a hypothesized pose explains the observed image.

- Phase 1 – Retrieval as the Prior

  We first use semantic mesh segmentation and viewport matching to rapidly narrow the search space from potentially thousands of poses to a small, high-recall set $\{\mathbf{P}_i\}_{i=1}^{K}$. The retrieval scores s_i assigned to these candidates are proportional to $p(\mathbf{P}_i \mid I_q, \mathcal{S})$ and thus act as an *empirical prior*. This step is deliberately tuned for coverage: even if some top-ranked candidates are wrong, the goal is to ensure that the true pose remains within the short list.

- Phase 2 – Descriptor Matching as the Likelihood

  We then introduce an *independent source* of evidence by computing rotation-invariant circular descriptors $\phi(\cdot)$ for both the query and each candidate rendering $\mathcal{R}(\mathbf{P}_i)$. The cosine similarity c_i between $\phi(I_q)$ and $\phi(\mathcal{R}(\mathbf{P}_i))$ measures how consistent the global layout is between the two views, regardless of in-plane rotation. We model this as a likelihood term:

  $$p(\mathbf{P}_i \mid I_q, \phi) \propto \exp(c_i), \tag{4}$$

  where the exponential ensures higher similarity values correspond to higher likelihood (by introducing soft-max style weighting that removes negative weights and magnifies stronger matches).
  The final posterior probability for each candidate pose is then:

  $$p(\mathbf{P}_i \mid I_q, \mathcal{S}, \phi) \propto p(\mathbf{P}_i \mid I_q, \mathcal{S}) \cdot p(\mathbf{P}_i \mid I_q, \phi). \tag{5}$$

 doi:10.20944/preprints202511.1027.v1

This Bayesian update ensures that candidates supported by *both* the high-recall retrieval prior and the ambiguity-resolving descriptor likelihood rise to the top. Importantly, it allows correct poses that were initially ranked low due to retrieval bias to overtake incorrect top-1 candidates. This is something correlation-based refinements cannot achieve reliably. In high-ambiguity layouts, this principled combination of evidence is the key to systematically improving accuracy without additional geometric solvers or per-scene retraining.

3.4. Rotation Invariance via Group Theory

A common error in retrieval-based localization is sensitivity to in-plane camera rotation (roll). Two images of the same location can yield very different descriptors if the representation is not rotation-invariant, which is especially problematic in panoramic or wide FoV settings where roll does not alter the scene layout.

Geometrically, the camera poses are in the Lie group $SE(3)$ of 3D rigid transformations. In-plane rotations form the subgroup $SO(2) \subset SE(3)$, representing rotations about the viewing axis. All poses in this $SO(2)$ subgroup are *orientation equivalent*, as they differ only in viewpoint orientation while preserving the same global layout.

Our rotation-invariant circular descriptor is designed to marginalize over this $SO(2)$ subgroup, effectively treating all in-plane rotations of the same scene as equivalent. In mathematical terms, if $g(\cdot)$ denotes the feature extractor applied to an image I, we approximate the marginalization

$$\phi(I) \approx \frac{1}{2\pi} \int_0^{2\pi} g(R_\theta I)\, d\theta, \tag{6}$$

where R_θ represents a rotation of the image by θ degrees. In practice, the feature map is transformed into polar coordinates and sampled using M concentric rings with N points each. Pooling (average and max) within every ring yields a compact descriptor of size $D = M \times N \times 2$, and concatenating across rings followed by L2 normalization produces the final rotation-invariant descriptor.

This construction (i) preserves global spatial cues such as the ordering of walls, doors, and structural boundaries while discarding arbitrary in-plane rotations, and (ii) enables direct cosine-based similarity between query and rendered views without orientation alignment. By embedding $SO(2)$ invariance into the descriptor, we remove a key nuisance factor from matching, which is especially beneficial in ambiguous layouts where disambiguation relies on large-scale geometry rather than orientation sensitive texture cues. Hence, the system is kept rotation-invariant to in-plane camera spins while still keeping the ability to reason about the larger 3D structure of the scene.

3.5. Algorithm Pipeline

- Stage 1: Initialization

  At query time, we begin by semantically segmenting the 3D room mesh and then render a set of panoramic viewports on a uniform 1.2 m × 1.2 m grid, requiring far fewer samples than LaLaLoc's [22] layout based approach. Inspired by the basal pipeline in [9], each rendered panorama and the input perspective image are passed through pretrained backbones: EfficientNet-S [37] for the query images and DenseNet [38] for the panoramas, to produce dense feature maps.

- Stage 2: Initial Pose Estimation

  We compute depth-wise correlations between the query's features and each panorama's features, yielding a similarity score for every candidate viewport. A lightweight MLP takes the top-k scoring candidates and directly regresses their 6-DoF poses. All training and evaluation follow the standard ZInD split [11]. This coarse retrieval stage (Figure 2 steps (a) and (b)) efficiently narrows the search from thousands of potential views to a few high-confidence hypotheses and does not require 2D floorplans or low-level annotations as in LASER [10].

 doi:10.20944/preprints202511.1027.v1

- Stage 3: Circular Feature Extraction

  We compute rotation-invariant circular descriptors for the query (and later, for each candidate viewport):

  1. Extract a dense feature map from a pretrained ResNet-50 [39] backbone.
  2. Transform into polar coordinates, sampling M rings and N points per ring ($M = 1$ and $N = 8$).
  3. Apply average and max pooling per ring to form D-dimensional vectors ($D = 256$).
  4. Concatenate across rings and L2-normalize.

  This stage corresponds to Figure 2 step (c) and produces a compact, rotation-agnostic signature that captures global layout, crucial for disambiguating visually similar scenes.

- Stage 4: Pose Re-ranking

  Finally, as can be seen in Figure 2 step (d), each candidate from *Stage 1* is re-rendered and encoded into a circular descriptor. Cosine similarity to the query descriptor is computed, and candidates are re-ranked. Unlike SPVLoc's correlation-based refinement, which assumes top-1 correctness, our Bayesian combination allows lower-ranked but correct candidates to surface. This avoids failure modes where initial retrieval bias dominates refinement.

4. Experimental Evaluation

4.1. Baselines

We evaluate Render-Rank-Refine against existing camera pose-estimation baselines (LASER [10] and SPVLoc [9]) on the Zillow Indoor Dataset (ZInD) split [11]. Quantitative performance is measured by translation error (T_{err}, cm) and rotation error (R_{err},°) related statistics over all test scenes. Qualitative examples illustrate success cases of our strategy in challenging rooms with overlapping layouts. The GitHub repository for this project is currently under development and will be released in due course. The repositories for SPVLoc and LASER are publicly available online for reference.

4.2. Dataset

We validate on the Zillow Indoor Dataset (ZInD) [11], which contains 67,448 panoramas from 1,575 unfurnished residential homes, each aligned with its floor plan. ZInD also encodes features such as open corridors connecting rooms without doors, introducing greater structural diversity in semantic renderings. Its rich 3D annotations allow direct generation of room-level mesh models, which can be converted into semantic meshes either through manual annotation of mesh faces or via learning-based segmentation methods such as PointNet++ [40] or graph neural networks for meshes [41]. Modern frameworks and consumer devices further enable rapid 3D mesh generation without manual annotations or floorplans, such as Open3D [42], LiDAR-equipped smartphones and ARKit-based apps (e.g., Polycam, SiteScape) [43].

4.3. Model Training

We initialize all network components with the official SPVLoc [9] pre-trained weights, leveraging their robust semantic and geometric feature representations for faster convergence. The model is fine-tuned once, end-to-end, on our dataset, which is a one-time process that does not require per-scene retraining and is therefore excluded from our computational burden analysis. To maximize hardware utilization, training uses a batch size of 40 and 24 CPU workers (our system's limit). All other training hyperparameters (optimizer, learning rate schedule, etc.) follow the original implementation to ensure consistency and reproducibility. Minor deviations in error metrics may arise from hardware differences and numerical precision changes (such as float32 instead of float64) necessary for compatibility.

 doi:10.20944/preprints202511.1027.v1

4.4. Inference Results.

Quantitative Evaluation Table 1 presents translation and rotation recall at various thresholds, evaluated on perspective queries with a 90° field of view (FoV). While our recall values closely match those of the refined SPVLoc baseline, minor differences at these coarse thresholds overlook critical variations in *per-case* accuracy. This highlights the need to explore where our rotation-invariant refinement benefits, particularly in challenging, high-ambiguity scenes (high $A(v)$, see Figure 5). Rather than focusing on aggregate recall, we break down the maximum, minimum, mean, and median T_{err} (cm) and R_{err} (∘) in Table 2 for a detailed analysis.

Table 1. Translation and rotational recall (%) at varying thresholds for different pipeline variants. *SPVLoc Refined* and *Ours Refined* refer to the baseline with one step of correlation-score-based refinement from [9].

Method	$<$ **10 cm** (%)	$<$ **50 cm** (%)	$<$ **1 m** (%)
LASER (from [9])	8.69	67.01	80.90
SPVLoc	12.25	59.49	98.21
SPVLoc – Refined	22.07	73.80	98.21
Ours	12.86	63.42	99.14
Ours – Refined	21.82	73.62	99.14

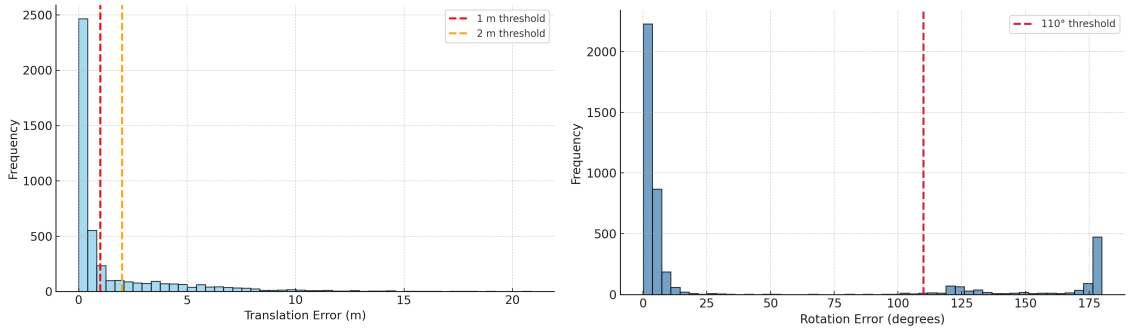
Table 2 shows the overall error statistics for all methods. The baseline method has the highest mean T_{err} and R_{err}, while refining the baseline reduces both errors. Our method further improves translation accuracy (mean T_{err} = 116.54 cm) while maintaining similar rotational accuracy. Notably, the circular feature-based reranking method delivers the lowest mean T_{err} and R_{err} and the smallest median T_{err}. Incorporating circular features improves localization accuracy, and further refinement (based on correlation as in [9]) improves performance beyond the baseline and the refined baseline.

Table 2. Overall translation (cm) and rotation (°) error statistics for all approaches.

	$T_{\mathbf{err}}$ **(cm)**				$R_{\mathbf{err}}$ **(°)**			
Method	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**	**Mean**	**Median**
Baseline	0.406	2089.24	149.93	32.38	0.040	179.99	38.91	3.57
Baseline – refined	1.210	2069.46	131.73	18.47	0.046	179.99	36.33	2.97
Ours	0.406	2089.24	116.54	29.17	0.040	179.99	35.98	3.44
Ours – refined	0.406	2069.47	103.07	19.41	0.027	179.99	30.20	2.56

Error distribution analysis. Figure 4 shows the distributions of translation error T_{err} (m) and rotation error R_{err} (°) of the baseline on the ZInD test set. To analyze performance, we use thresholds of 1 m and 2 m: the first captures near-accurate cases, while the second brackets the long-tail failures seen in the original pipeline. For rotation, we choose 110° using the same criteria. Most camera poses are predicted within 1 m, but a smaller number produce errors beyond 2 m. Rotation errors are mostly concentrated at small angles, with some outliers corresponding to large orientation flips.

 doi:10.20944/preprints202511.1027.v1



(a) (b)

Figure 4. Error distributions on ZInD: (**a**) Translation error distribution with 1 m and 2 m thresholds. Most T_{err} values lie below 1 m with a long tail beyond 2 m. (**b**) Rotation error distribution with a 110° threshold. R_{err} concentrates at small angles, with a secondary mass near extreme flips. Thresholds are chosen from the worst observed T_{err} and a practical failure bound of 110° for rotation.

To probe the causes of large errors, we randomly sampled 20 queries beyond each threshold. For $T_{err} > 1$ m, 14 of the 20 (70%) images came from overlapping layouts; for $T_{err} > 2$ m, 13 out of 20 (65%) were overlapping. For severe rotation outliers with $R_{err} > 110°$, 11/20 (55%) of the sampled images also originated from overlapping layouts. All failure cases point back to structural ambiguity (high $A(v)$) that can be seen in query images covering more than one room, motivating layout-level disambiguation (e.g., multi-view reasoning or explicit layout priors) in future work. Sampling another 20 queries randomly produced approximately identical results.

Reverse validation. To further prove our method's superior accuracy on query scenarios with high ambiguity, we analyzed the ten scenes with the highest translation errors in the baseline method and examined their corresponding rotation and translation errors. Nearly all of these challenging cases were also found to resemble the ambiguous room layouts as in Figure 1. The mean and median percentage improvements across the hardest cases (top-10 worst T_{err}) are substantial. Compared to the refined baseline, which yields only marginal gains ($< 1\%$ mean improvement in translation and $\sim$12% in rotation, with virtually no cases above 90% translation improvement), our method achieves far larger gains: a 40.4% mean and 5.2% median reduction in T_{err}, alongside a 29.7% mean reduction in R_{err}. Notably, 4/10 of these hardest cases reach at least 90% improvement in translation error and 30% achieve this for rotation error, underscoring the robustness of our approach in highly ambiguous settings. On the remaining 6/10 cases, the improvement was found to be more than 50% in both errors. This represents a substantial improvement over the refined baseline (best-case baseline) which achieves such mean gains in rotation error for only 1/10 of the cases and fails to do so in translation accuracy.

Qualitative Analysis. In addition to the ambiguous scenes shown in Figure 1, we further examine challenging indoor environments that frequently degrade pose estimation accuracy. Figure 5 shows some representative examples from the ZInD dataset on which our method outperforms the existing state-of-the-art methods of camera pose estimation:

(a) Multi-room View through Cutouts: Overlapping rooms connected by partial wall openings introduce mixed semantic cues. Our approach successfully isolates the intended viewport by comparing rendered perspectives against the query in a rotation-invariant feature space.

(b) Kitchen with Foreground Occlusion: Complex indoor environments containing strong occlusions (e.g.,, light fixtures) can obscure discriminative features. The circular descriptors retain robustness to such occlusions, enabling correct disambiguation from visually similar kitchen layouts.

(c) Transitional Hallway - i: Areas connecting multiple rooms present high ambiguity due to overlapping layouts and similar lighting conditions. Our reranking step leverages viewpoint-independent features to identify the correct spatial configuration and reject misleading candidates.

(d) Transitional hallway - ii: The repeated door frames and partial room glimpses create visually similar, overlapping layout cues and weak semantics.

 doi:10.20944/preprints202511.1027.v1

Figure 5. Qualitative results on challenging ZInD [11] scenes. Each illustrates a visually ambiguous layout where baseline methods fail, but our method resolves the camera pose using rotation-invariant circular descriptors. (**a**) Multi-room view. (**b**) Kitchen with occlusion. (**c**) Transitional hallway I. (**d**) Transitional hallway II.

Runtime Analysis. All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 2080 Ti (TU102 Rev. A) GPU with 11 GB VRAM. Nonetheless, all computations used CPU-based multi-threaded processing rather than GPU acceleration. Table 3 compares the runtime characteristics of our method against existing approaches, including PFNet [44] and LASER [10].

Table 3. Runtime comparison of scene sampling time (mean ± std, seconds) and inference throughput (queries per second, QPS).

Method	Scene sampling time (s)	Inference QPS
PFNet [44]	48.95 ± 38.95	5.06
LASER [10]	0.97 ± 1.09	8.31
Baseline [9]	1.66 ± 1.06	**28.63**
Ours	$\mathbf{0.72 \pm 0.65}$	**25.79**
Ours (refined)	$\mathbf{0.70 \pm 0.63}$	**26.39**

Our approach achieves a mean scene sampling time of 0.717 s without refinement and 0.704 s with refinement, representing a $\sim$5.7 and $\sim$5.8 times speedup over the baseline, without and with refinement, respectively. In terms of inference throughput, our method sustains 25.8–26.8 QPS (queries per second), which is 4 times faster than LASER (5.5 QPS) and more than 12 times faster than PFNet (2.2 QPS), while maintaining state-of-the-art localization accuracy. QPS is an estimate of how many query images the pipeline can localize per second at runtime, end-to-end.

These results show that circular-descriptor reranking adds little overhead while sharply reducing sampling time. Using lightweight features and cosine similarity instead of heavy inference, it supports near real-time deployment in large indoor environments without loss of accuracy.

 doi:10.20944/preprints202511.1027.v1

5. Discussion and Limitation

Our rotation-invariant circular reranking improves accuracy and efficiency across settings. Relative to the baseline [9], it cuts mean scene-sampling time by 5.7–5.8 times and sustains 25.8–26.8 QPS (about 4 times LASER [10] and greater than 12 times PFNet [44]), with negligible overhead thanks to lightweight feature extraction and cosine matching (no dense latent rendering). The method also remains reliable in challenging layouts. In ZInD, many queries span overlapping or multi-room views where a single FoV covers distinct spaces that confound viewpoint-specific descriptors. Our rotation-invariant descriptors encode global layout, enabling disambiguation even when local semantics are weak or misleading.

From a Bayesian view, SPVLoc retrieval provides a high-recall but orientation-sensitive prior, while circular descriptors supply a rotation-invariant likelihood; the posterior promotes correct poses without assuming the top-1 candidate, addressing the retrieval-bias failure mode noted in Section 1. Group theoretically, treating in-plane rotations as $SO(2) \subset SE(3)$ and marginalizing them via polar sampling yields descriptors that preserve global layout and perform well in symmetric or repetitive scenes. The framework is modality agnostic (such as RGB-LiDAR), requires neither floor plans nor per-scene retraining, and suits latency-sensitive smart health, AR, and robotics. Future work will improve robustness to illumination and evaluate app-generated semantics with native perspective images.

Our method occasionally underperforms in visually uniform or low-texture scenes (e.g., plain corridors or doorways with repetitive geometry). In such cases, the rotation-invariant reranking may struggle to converge, as the absence of distinctive gradients or depth cues leads to ambiguous feature alignment and local minima.

Figure 6 illustrates one such instance where flat, symmetric surfaces reduce descriptor discriminability, degrading pose precision despite an otherwise stable pipeline. In some other cases, vertical or multi-floor ambiguity may persist due to the absence of reliable height cues. Further, our evaluation is currently limited to residential datasets (like ZInD), so generalization to other building types (such as hospitals and industrial) and to native perspective captures remains to be validated.

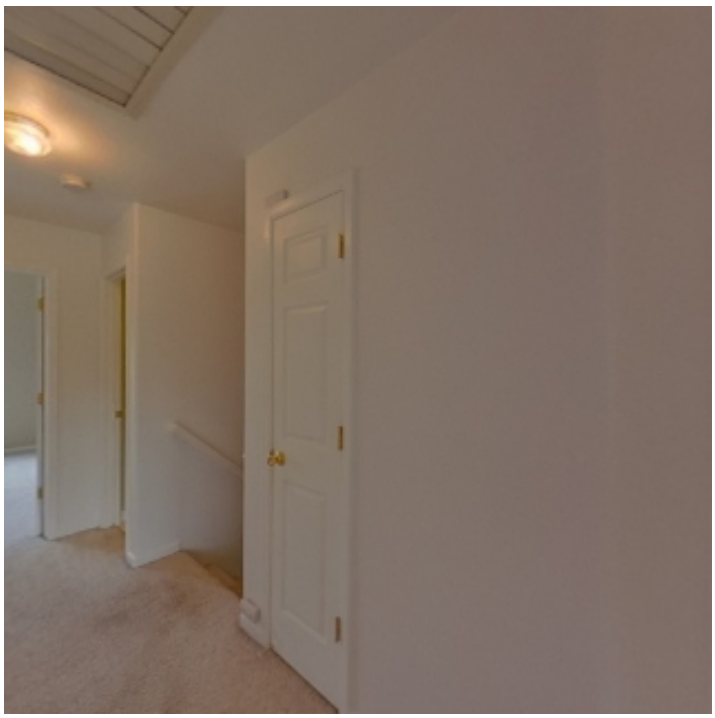


Figure 6. Example failure case from ZInd. A low-texture hallway where repetitive geometry and weak visual cues cause refinement to worsen relative to the baseline.

 doi:10.20944/preprints202511.1027.v1

6. Conclusions

We present Render-Rank-Refine, a two-stage framework for robust indoor camera pose estimation. The approach treats candidate retrieval as a high-recall prior and circular descriptor matching as an ambiguity-resolving likelihood, making the final refinement a coherent probabilistic update. From a group-theoretic perspective, marginalizing in-plane rotations removes a nuisance variable while preserving the global spatial layout—thereby reducing retrieval bias and stabilizing pose estimates in repetitive environments. On the Zillow Indoor Dataset, our method achieves substantial speedups over prior approaches while maintaining or improving accuracy, with the most significant gains observed in ambiguous layouts. The framework is efficient, lightweight, and modality-agnostic, naturally extending beyond RGB to cross-domain scenarios, making it well suited for real-time applications in AR, robotics, and assistive systems.

Acknowledgments: During the preparation of this manuscript/study, the author(s) used ChatGPT-5 for the purposes of correcting text grammatically and correcting python codes where needed. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

References

1. Taira, H.; Okutomi, M.; Sattler, T.; Cimpoi, M.; Pollefeys, M.; Sivic, J.; Pajdla, T.; Torii, A. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. In Proceedings of the Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, [arXiv:cs.CV/1803.10368].
2. Wang, S.; Fidler, S.; Urtasun, R. Lost Shopping! Monocular Localization in Large Indoor Spaces. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015.
3. Acharya, D.; Ramezani, M.; Khoshelham, K.; Winter, S. BIM-Tracker: A model-based visual tracking approach for indoor localisation using a 3D building model. *ISPRS Journal of Photogrammetry and Remote Sensing* **2019**, *150*, 157–171. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2019.02.014.
4. Chen, X.; Fan, G.; Roberts, E.; Steven, H.J. A Transfer Learning-Based Smart Homecare Assistive Technology to Support Activities of Daily Living for People with Mild Dementia. In Proceedings of the 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), 2023, pp. 359–363. https://doi.org/10.1109/BIBE60311.2023.00065.
5. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle adjustment—a modern synthesis. In Proceedings of the International workshop on vision algorithms. Springer, 1999, pp. 298–372.
6. Snavely, N.; Seitz, S.; Szeliski, R. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics (2006)* **2006**, *25*, 835–846.
7. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization, 2016, [arXiv:cs.CV/1505.07427].
8. Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; Zhou, Z. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. In Proceedings of the Proceedings of The European Conference on Computer Vision (ECCV), 2020.
9. Gard, N.; Hilsmann, A.; Eisert, P. SPVLoc: Semantic Panoramic Viewport Matching for 6D Camera Localization in Unseen Environments. In Proceedings of the Computer Vision – ECCV 2024: 18th European Conference on Computer Vision, Proceedings, Part LXXIII, Cham, 2024; pp. 398–415. https://doi.org/10.1007/978-3-031-73464-9_24.
10. Min, Z.; Khosravan, N.; Bessinger, Z.; Narayana, M.; Kang, S.B.; Dunn, E.; Boyadzhiev, I. LASER: LAtent SpacE Rendering for 2D Visual Localization, 2023, [arXiv:cs.CV/2204.00157].
11. Cruz, S.; Hutchcroft, W.; Li, Y.; Khosravan, N.; Boyadzhiev, I.; Kang, S.B. Zillow Indoor Dataset: Annotated Floor Plans With 360º Panoramas and 3D Room Layouts. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 2133–2143.
12. Torii, A.; Sivic, J.; Pajdla, T.; Okutomi, M. Visual Place Recognition with Repetitive Structures. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 883–890. https://doi.org/10.1109/CVPR.2013.119.
13. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14136–14147. https://doi.org/10.1109/CVPR46437.2021.01392.

 doi:10.20944/preprints202511.1027.v1

14. Sarlin, P.E.; Cadena, C.; Siegwart, R.; Dymczyk, M. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12708–12717. https://doi.org/10.1109/CVPR.2019.01300.
15. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **2004**, *60*, 91–110.
16. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* **2008**, *110*, 346–359. Similarity Matching in Computer Vision and Multimedia, https://doi.org/https://doi.org/10.1016/j.cviu.2007.09.014.
17. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, 2011, pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544.
18. Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary Robust invariant scalable keypoints. In Proceedings of the 2011 International Conference on Computer Vision, 2011, pp. 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542.
19. Luo, Z.; Zhou, L.; Bai, X.; Chen, H.; Zhang, J.; Yao, Y.; Li, S.; Fang, T.; Quan, L. ASLFeat: Learning Local Features of Accurate Shape and Localization. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6588–6597. https://doi.org/10.1109/CVPR42600.2020.00662.
20. Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In *Readings in Computer Vision*; Fischler, M.A.; Firschein, O., Eds.; Morgan Kaufmann: San Francisco (CA), 1987; pp. 726–740. https://doi.org/https://doi.org/10.1016/B978-0-08-051581-6.50070-2.
21. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EP n P: An accurate O (n) solution to the P n P problem. *International journal of computer vision* **2009**, *81*, 155–166.
22. Howard-Jenkins, H.; Ruiz-Sarmiento, J.R.; Prisacariu, V.A. Lalaloc: Latent layout localisation in dynamic, unvisited environments. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10107–10116.
23. Howard-Jenkins, H.; Prisacariu, V.A. Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 693–709.
24. Salas-Moreno, R.F.; Newcombe, R.A.; Strasdat, H.; Kelly, P.H.; Davison, A.J. Slam++: Simultaneous localisation and mapping at the level of objects. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 1352–1359.
25. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In Proceedings of the European conference on computer vision. Springer, 2014, pp. 834–849.
26. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics* **2015**, *31*, 1147–1163. https://doi.org/10.1109/TRO.2015.2463671.
27. Mur-Artal, R.; Tardos, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics* **2017**, *33*, 1255–1262. https://doi.org/10.1109/tro.2017.2705103.
28. Whelan, T.; Salas-Moreno, R.F.; Glocker, B.; Davison, A.J.; Leutenegger, S. ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research* **2016**, *35*, 1697–1716.
29. McCormac, J.; Handa, A.; Davison, A.; Leutenegger, S. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 4628–4635. https://doi.org/10.1109/ICRA.2017.7989538.
30. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, 2020, [arXiv:cs.CV/2003.08934].
31. Dai, A.; Ritchie, D.; Bokeloh, M.; Reed, S.; Sturm, J.; Nießner, M. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4578–4587. https://doi.org/10.1109/CVPR.2018.00481.
32. Brahmbhatt, S.; Gu, J.; Bansal, K.; Darrell, T.; Hwang, J.; Adelson, E.H. Geometry-Aware Learning of Maps for Camera Localization. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
33. Clark, R.; Wang, S.; Wen, H.; Handa, A.; Nieuwenhuis, D.; Davison, A.; Leutenegger, S. VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

 doi:10.20944/preprints202511.1027.v1

34. Tateno, K.; Tombari, F.; Laina, I.; Navab, N. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6243–6252.
35. Brachmann, E.; Krull, A.; Nowozin, S.; Shotton, J.; Michel, F.; Gumhold, S.; Rother, C. DSAC — Differentiable RANSAC for Camera Localization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2492–2500. https://doi.org/10.1109/CVPR.2017.267.
36. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014, pp. 15–22. https://doi.org/10.1109/ICRA.2014.6906584.
37. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020, [arXiv:cs.LG/1905.11946].
38. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks, 2018, [arXiv:cs.CV/1608.06993].
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
40. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *CoRR* **2017**, *abs/1706.02413*, [1706.02413].
41. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *CoRR* **2018**, *abs/1801.07829*, [1801.07829].
42. Zhou, Q.Y.; Park, J.; Koltun, V. Open3D: A Modern Library for 3D Data Processing, 2018, [arXiv:cs.CV/1801.09847].
43. Askar, C.; Sternberg, H. Use of Smartphone Lidar Technology for Low-Cost 3D Building Documentation with iPhone 13 Pro: A Comparative Analysis of Mobile Scanning Applications. *Geomatics* **2023**, *3*, 563–579. https://doi.org/10.3390/geomatics3040030.
44. Huang, Z.; Yu, Y.; Xu, J.; Ni, F.; Le, X. PF-Net: Point Fractal Network for 3D Point Cloud Completion, 2020, [arXiv:cs.CV/2003.00410].