ActLoc: Learning to Localize on the Move via Active Viewpoint Selection

Jiajie Li[1,*], Boyang Sun[1,*], Luca Di Giammarino[2], Hermann Blum[1,4], Marc Pollefeys[1,3]
[1]ETH Zürich [2]Sapienza University of Rome [3]Microsoft [4]University of Bonn
[*]Equal contribution https://boysun045.github.io/ActLoc-Project/

***Abstract*—Reliable localization is critical for robot navigation, yet most existing systems implicitly assume that all viewing directions at a location are equally informative. In practice, localization becomes unreliable when the robot observes unmapped, ambiguous, or uninformative regions. To address this, we present ActLoc, an active viewpoint-aware planning framework for enhancing localization accuracy for general robot navigation tasks. At its core, ActLoc employs an attention-based model that encodes a metric map and the camera poses used during map construction, and predicts localization accuracy across yaw and pitch directions at arbitrary 3D locations in a single forward pass. These per-point accuracy distributions are incorporated into a path planner, enabling the robot to actively select camera orientations that maximize localization robustness while respecting task and motion constraints. ActLoc achieves state-of-the-art results on single-viewpoint selection and generalizes effectively to full-trajectory planning.**

I. Introduction

Visual localization has become increasingly important in robotics due to the affordability and versatility of cameras compared to LiDAR sensors. Visual maps can also be shared across devices, enabling multi-agent spatial alignment [1], [2] and facilitating human–robot collaboration [3]. However, camera-based localization remains more challenging than LiDAR-based positioning [4]. A key limitation stems from the narrow field of view (FoV) of cameras, which prevents omni-directional sensing. Since visual localization relies heavily on image features, viewpoints covering feature-rich regions tend to yield more reliable results. This observation has motivated active localization strategies [5]–[7], where viewpoint selection is integrated into the robot's motion planning to improve localization accuracy along the trajectory.

However, existing active localization methods face several limitations. Learning-based approaches improve accuracy but remain computationally expensive, since evaluating localization quality at each viewpoint demands substantial resources. Moreover, most methods assess viewpoints independently, requiring sparse sampling and decoupling viewpoint selection from the path planning process.

To address these challenges, we propose **ActLoc**, an efficient and scalable active localization framework. Instead of evaluating viewpoints individually, our attention-based model encodes scene information and directly predicts a distribution of localization quality across multiple yaw–pitch directions at each 3D waypoint in a single forward pass. This enables trajectory-level optimization that jointly accounts for task objectives and localization robustness, while being computationally efficient for real-time deployment.

Our main contributions are:

- A learning-based model that predicts a distribution of localization accuracy over multiple yaw–pitch directions in a single forward pass.
- **ActLoc**, an active localization planner that integrates the model into a navigation system for online viewpoint selection and trajectory planning.
- Comprehensive experiments showing that ActLoc consistently outperforms baselines in both single-viewpoint selection and long-horizon path planning.

II. Related Work

Visual Localization. Visual localization estimates the 6-DoF camera pose of a query image within a known environment. Approaches include two-step methods that establish correspondences between query and reference images via sparse [8]–[11] or dense features [12], [13], and direct pose regression methods [14]–[16]. While both paradigms have achieved strong results, they remain inherently *passive*: the camera pose is inferred solely from the captured view, without reasoning about which viewpoint would yield the most reliable localization.

Active Viewpoint Selection. Active visual localization extends beyond passive pose estimation by allowing an agent to actively select viewpoints to improve localization reliability. This direction builds upon active perception [17]–[19], which studies strategies for guiding sensor orientations and robot motion to maximize the informativeness of observations. Early studies focused on geometric metrics such as Fisher Information [20] to identify informative viewpoints, but such hand-crafted criteria are environment-specific and computationally expensive for dense sampling. Recent learning-based methods span pose regression [21], policy learning [22], [23], and viewpoint classification [6], but often rely on reinforcement learning or user-annotated supervision that limits generalization. The most relevant work, LWL [7], trains a lightweight MLP to classify viewpoints as suitable or unsuitable using self-supervised labels from voxel-grid sampling and visibility checks. However, it requires dense viewpoint sampling and extensive preprocessing at inference, making it computationally costly. Furthermore, while LWL presents qualitative planning results, it lacks a fully integrated and quantitatively evaluated

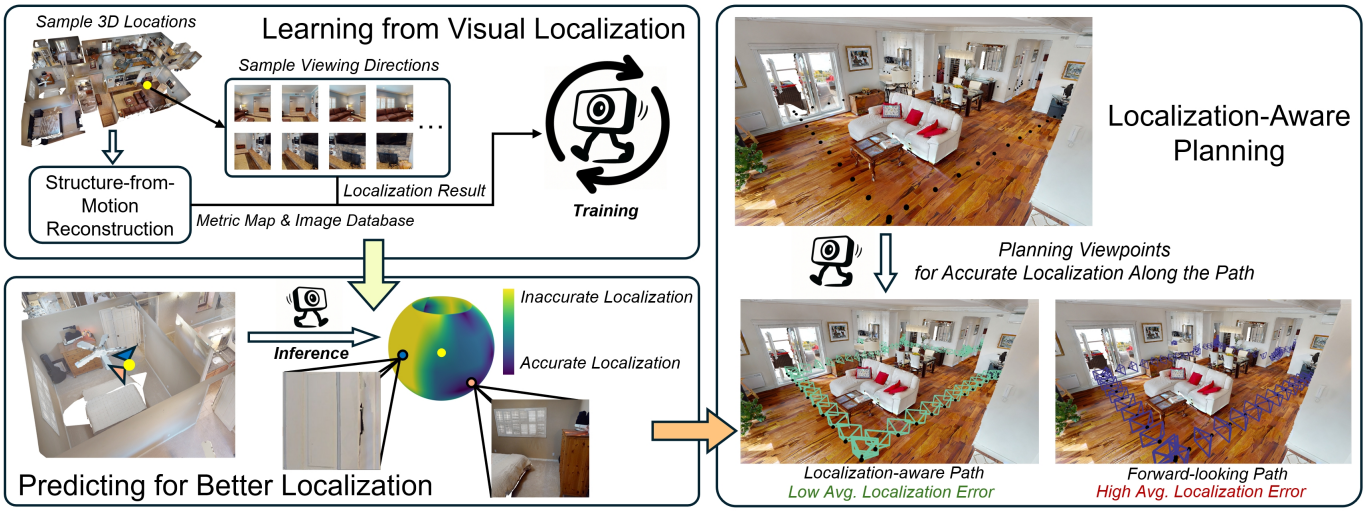

Fig. 1. **ActLoc** is an active viewpoint selection framework designed to improve visual localization for robot navigation. **Left**: The model is trained using supervision from Visual Localization, learning to predict distributions of localization accuracy over yaw and pitch angles at arbitrary 3D locations. **Right**: The trained model is integrated into the path planning process, enabling the robot to select viewpoints at each waypoint that increase localization robustness along its trajectory.

trajectory-level planning system. Our approach removes these limitations by predicting localization performance across all viewpoints in a single pass and integrating these predictions into motion planning.

III. Method

A. System Overview

Our system consists of two main components: (1) an attention-based model trained on Structure-from-Motion (SfM) data that predicts visual localization performance over local yaw and pitch angles in a single forward pass, which we denote as *LocMap*; and (2) the integration of LocMap into a path planning pipeline, which, given a sequence of *waypoints* (robot positions in 3D space), actively selects *viewpoints* (camera orientations at each waypoint) to improve localization accuracy while respecting motion constraints. Figure 2 provides an overview.

B. Predicting Localization Performance

At each waypoint, the planner must choose a camera orientation that yields reliable localization. We assume the environment has been previously mapped, which is standard in visual localization since methods typically query an existing database. This assumption naturally applies to many robotics tasks, such as inspection, monitoring, and indoor service. Our model encodes prior scene knowledge and predicts the expected localization performance across multiple viewing directions. The output is the *LocMap*, a grid of size $H \times W$, where each cell represents the predicted localization performance for a candidate viewpoint defined by its yaw (horizontal) and pitch (vertical) angles. We exclude extreme pitch angles that point toward the ceiling or floor, as well as in-plane rotations around the camera's principal axis, since these provide negligible benefit for visual localization [24].

Unlike prior work that either scores only the best viewpoint [20] or evaluates multiple viewpoints independently without modeling correlations across orientations [6], [7], our model predicts a distribution of localization performance over all candidate viewpoints in a single forward pass. This design reflects the natural fact that multiple viewpoints may provide good localization, while also avoiding redundant computation incurred by per-viewpoint inference. Moreover, modeling the joint distribution allows the network to capture spatial correlations between viewpoints, which is crucial for accurately predicting localization performance at a waypoint.

We propose an attention-based model whose inputs come from the SfM reconstruction database, which typically maintains both a metric map and the camera poses used for reconstruction. Given the 3D landmarks $\mathbf{P} \in \mathbb{R}^{3\times N}$ and camera poses $\mathbf{T} \in \mathbb{SE}(3)^M$ from this database, and a waypoint $\mathbf{x} \in \mathbb{R}^3$, the model predicts:

$$\mathbf{C}_{\mathbf{x}}^{H\times W} = f_\theta(\mathbf{P}, \mathbf{T}, \mathbf{x}) \tag{1}$$

where each cell stores the likelihood that the localization error at the corresponding yaw–pitch direction falls below a threshold.

C. Model Architecture

The model takes multi-modal input from landmarks and camera poses. Given the raw inputs, we first remove highly uncertain 3D landmarks from $\mathbf{P}$, following [7]. Next, we apply an egocentric transformation that shifts $\mathbf{T}$ and the filtered $\mathbf{P}$ from the world frame to the local frame of $\mathbf{x}$. Importantly, we shift only their origins while preserving rotations to maintain consistency with the default viewing direction at $\mathbf{x}$. This implicitly incorporates waypoint location into the inputs without additional modules. Finally, we crop landmarks within a local

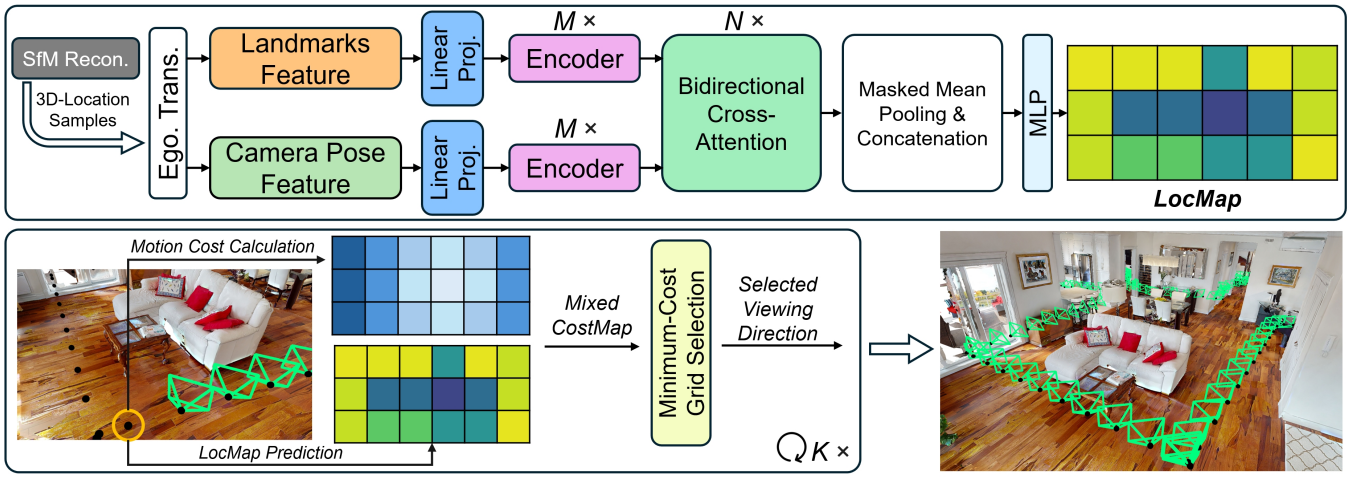

Fig. 2. **System Overview. Top**: The core of ActLoc is an attention-based network that processes multi-modal inputs from an SfM model, including the metric map and camera poses. The network predicts localization performance as a two-dimensional grid over multiple pitch and yaw rotations. **Bottom**: During path planning, ActLoc constructs a mixed cost map at each candidate location by combining the network prediction with additional factors, such as motion constraints, and selects the viewing direction that minimizes this cost.

bounding box centered at $\mathbf{x}$ to enforce locality and normalize input scale.

Camera features are represented as 7D vectors $(\mathbf{p_c}, \mathbf{q_c})$ (position and quaternion), and landmark features as 6D vectors (3D coordinates and normalized RGB). Both are projected into a shared embedding space using linear layers, then processed by transformer encoder blocks to extract context-aware representations. A bidirectional cross-attention module fuses the two modalities, allowing pose embeddings to attend to landmarks and vice versa. Since attention operates on sequences, we apply masked mean pooling (ignoring padding) to obtain one fixed-size vector per modality; these are concatenated and fed to an MLP to produce the final LocMap logits.

D. Training Data

We train on simulated data from HM3D [25], using 100 single-floor scenes (90 train/val, 10 test). To construct prior maps, we sample free-space points at heights between 1.5 and 2.0 meters, where the camera performs a full 360° sweep in 36° azimuth increments with random elevation angles from $[-15^\circ, 15^\circ)$. SfM is then run on these images using SuperPoint [24] features via hloc [9] to obtain 3D landmarks.

Training waypoints are independently sampled at heights between 0.4 and 2.0 meters, making our approach free of grid-based assumptions as in [7]. At each waypoint, the camera rotates at 20° intervals covering the full yaw range, with elevation varied from $[-60^\circ, 60^\circ)$ at 20° intervals. Each viewpoint is localized against the metric map using hloc [9] and COLMAP [26], yielding ground-truth localization performance labels automatically.

E. Path Planning with Localization Awareness

LocMap provides a structured representation for viewpoint selection in both single- and multi-viewpoint planning. Unlike prior work that evaluates viewpoints independently, LocMap predicts the distribution of localization performance across all candidate viewpoints at a waypoint in a single inference. This enables optimizing viewpoint choices along a waypoint sequence for higher overall localization accuracy.

For a given waypoint $\mathbf{x}$, the LocMap $\mathbf{C_x}$ provides a direct selection criterion: choosing the grid cell with the highest predicted localization performance. However, during path execution the robot must also satisfy motion constraints, such as limiting rotations between consecutive frames. We define a mixed cost map:

$$\mathbf{M_x} = \mathbf{C_x} + \lambda \cdot \mathbf{D_x} \tag{2}$$

where $\mathbf{D_x}$ is the Mahalanobis distance from each candidate viewpoint to the previously selected viewpoint, and λ balances localization accuracy and smooth motion. The robot selects the viewpoint with the minimum cost at each step. This strategy can be extended to incorporate additional task-specific metrics.

IV. Experiments

We evaluate ActLoc on two tasks: (a) single-viewpoint selection on HM3D and ScanNet scenes, comparing against FIF and LWL; and (b) trajectory-level path planning on five HM3D scenes, comparing against LWL and a heuristic forward-facing baseline. **FIF** [20] is an information-theoretic method based on Fisher Information that estimates localization quality via Gaussian Process regression and multiple information-theoretic metrics (minimum eigenvalue, determinant, trace), where we report the best-performing metric. **LWL** [7] is a learning-based method that classifies viewpoint suitability using visibility checks and image binning, trained on the same dataset as ours for fairness. All experiments use a single NVIDIA RTX 4090 GPU.

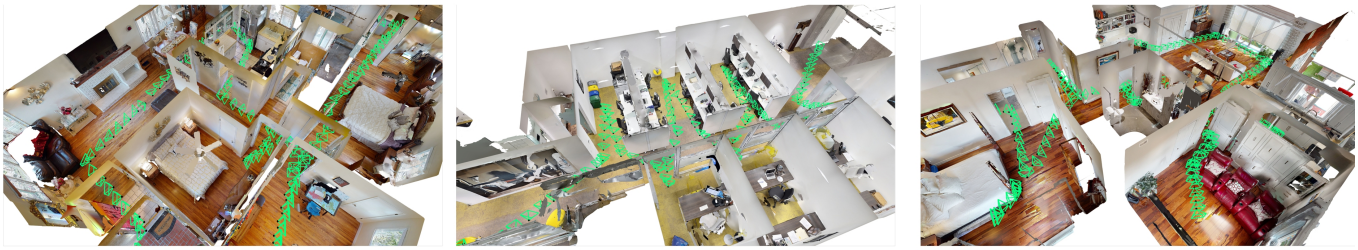

Fig. 3. **Qualitative Planning Results.** Visualization of ActLoc's viewpoint selection along waypoint sequences across different test scenes. The planned viewpoints demonstrate how ActLoc adapts camera orientation dynamically to balance localization performance and motion smoothness.

	Method	0.1m/1°	0.25m/2°	0.5m/5°	5m/10°
HM3D	LWL [7]	90.04	90.98	91.17	91.92
	FIF [20]	60.71	62.41	63.53	64.66
	ActLoc	**92.11**	**92.48**	**92.86**	**93.23**
	Upper Bound	97.74	97.74	97.93	98.50
Scannet	LWL [7]	88.89	89.90	89.90	92.93
	ActLoc	**91.92**	**94.95**	**96.97**	**96.97**
	Upper Bound	100.00	100.00	100.00	100.00

TABLE I

SINGLE-VIEWPOINT SELECTION RESULTS. LOCALIZATION SUCCESS RATES (%) ON TEN HM3D AND FIVE SCANNET SCENES. METRICS FOLLOW THE LONG-TERM VISUAL LOCALIZATION BENCHMARK [28], WITH AN ADDITIONAL FINE-GRAINED INDOOR THRESHOLD (SIMILAR TO [7]). ACTLOC ACHIEVES THE BEST RESULTS ACROSS ALL THRESHOLDS.

Scene ID	Forward	LWL [7]	ActLoc
00005	40.26 / 0.09 / 1.21	36.25 / 0.10 / 1.20	**52.75** / 0.07 / 1.09
00080	11.03 / 0.19 / 1.17	3.23 / 0.24 / 1.57	**24.00** / 0.05 / 0.80
00105	20.10 / 0.22 / 2.34	10.54 / 1.13 / 3.78	**27.25** / 0.16 / 1.92
00110	12.45 / 0.20 / 2.13	31.40 / 0.09 / 1.71	**35.95** / 0.07 / 1.37
00254	29.61 / 0.13 / 1.27	27.11 / 0.08 / 1.43	**46.08** / 0.14 / 1.62

TABLE II

PATH PLANNING RESULTS. *Success Rate (%) / Trans. Err. (m) / Rot. Err. (°)* ACROSS FIVE HM3D TEST SCENES. ACTLOC CONSISTENTLY ACHIEVES THE HIGHEST SUCCESS RATE.

A. Single-viewpoint Selection

Following the protocol of [7], we report the percentage of waypoints localized within four thresholds. We also report a theoretical upper bound, counting a waypoint as successfully localized if *any* of its candidate viewpoints meets the threshold. Our model outputs a 6×18 LocMap, adopted as a trade-off between efficiency and accuracy, and selects the viewpoint with the highest predicted localization probability. For a fair comparison, LWL is evaluated over the same 6×18 set of viewpoints, with the best-scoring direction chosen.

As shown in Table I, ActLoc consistently outperforms both baselines at all thresholds on both HM3D and ScanNet [27]. Notably, our model predicts localization performance across all candidate viewpoints simultaneously—a harder task than LWL's single-viewpoint formulation—yet achieves higher accuracy. These results highlight the model's ability to capture global spatial structure from SfM inputs and to generalize robustly to unseen scenes. The average processing time per waypoint is **110 ms** for ActLoc versus **8230 ms** for LWL, a $\sim$75$\times$ speedup. This efficiency gap grows further with higher-resolution LocMaps or larger-scale scenes.

B. Path Planning Evaluation

We evaluate ActLoc in a planning setting on five HM3D scenes with sparsified reconstructions to approximate real-world conditions. Path planning is performed with OMPL [29] to obtain waypoint sequences, which are then interpolated to generate dense waypoints at 20 cm intervals. Viewpoint selection is compared across ActLoc, LWL, and a forward-facing baseline. For ActLoc, we tune λ in the mixed cost map (Equation (2)), and we restrict sampling in LWL to a local neighborhood around the previously selected viewpoint. Processing times are equalized between ActLoc and LWL for fairness.

We evaluate using two metrics: success rate and average error across successful localizations. A localization is considered successful if the error is below 5 m / 10°. Results are summarized in Table II. ActLoc achieves the highest success rate across all scenes and, in most cases, the lowest average errors. Importantly, since LWL requires per-frame inference, the equalized time budget forces it to sample fewer candidate viewpoints per frame, resulting in degraded performance. Figure 3 shows qualitative examples of viewpoint selection along waypoint sequences. Thanks to the mixed cost map, ActLoc not only improves localization performance but also maintains smooth rotations between consecutive frames, respecting the robot's kinematic constraints.

V. CONCLUSION

We presented ActLoc, an active viewpoint-aware planning framework that predicts localization performance across all candidate viewing directions at a waypoint in a single forward pass, achieving state-of-the-art accuracy on single-viewpoint selection with a $\sim$75$\times$ speedup and generalizing effectively to full-trajectory planning. By integrating these predictions into path planning, ActLoc enables robots to dynamically select camera orientations that maximize localization robustness while respecting motion constraints. Future work includes extending ActLoc to diverse feature types and dynamic environments.

REFERENCES

[1] A. Cramariuc, L. Bernreiter, F. Tschopp, M. Fehr, V. Reijgwart, J. Nieto, R. Siegwart, and C. Cadena, "maplab 2.0 – A Modular and Multi-Modal Mapping Framework," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 520–527, 2023.

[2] Y. Chang, Y. Tian, J. P. How, and L. Carlone, "Kimera-multi: a system for distributed multi-robot metric-semantic simultaneous localization and mapping," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 11 210–11 218.

[3] J. Chen, B. Sun, M. Pollefeys, and H. Blum, "A 3d mixed reality interface for human-robot teaming," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 327–11 333.

[4] P. Dellenbach, J.-E. Deschaud, B. Jacquet, and F. Goulette, "Ct-icp: Real-time elastic lidar odometry with loop closure." IEEE, 2022, pp. 5580–5586.

[5] J. Lim, N. Lawrance, F. Achermann, T. Stastny, R. Bähnemann, and R. Siegwart, "Fisher information based active planning for aerial photogrammetry." IEEE, 2023, pp. 1249–1255.

[6] M. Hanlon, B. Sun, M. Pollefeys, and H. Blum, "Active visual localization for multi-agent collaboration: A data-driven approach." IEEE, 2024.

[7] L. Di Giammarino, B. Sun, G. Grisetti, M. Pollefeys, H. Blum, and D. Barath, "Learning where to look: Self-supervised viewpoint selection for active localization using geometrical information." Springer, 2024, pp. 188–205.

[8] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," vol. 39, no. 9, pp. 1744–1756, 2016.

[9] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," 2019, pp. 12 716–12 725.

[10] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," 2021, pp. 3247–3257.

[11] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," 2019, pp. 8092–8101.

[12] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," 2018, pp. 4654–4662.

[13] S. Tang, C. Tang, R. Huang, S. Zhu, and P. Tan, "Learning camera localization via dense scene matching," 2021, pp. 1831–1841.

[14] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," 2019, pp. 3302–3312.

[15] Y. Shavit, R. Ferens, and Y. Keller, "Learning multi-scene absolute pose regression with transformers," 2021, pp. 2733–2742.

[16] S. Chen, X. Li, Z. Wang, and V. A. Prisacariu, "Dfnet: Enhance absolute pose regression with direct feature matching." Springer, 2022, pp. 1–17.

[17] N. Roy, W. Burgard, D. Fox, and S. Thrun, "Coastal navigation-mobile robot navigation with uncertainty in dynamic environments," vol. 1. IEEE, 1999, pp. 35–40.

[18] G. Costante, C. Forster, J. Delmerico, P. Valigi, and D. Scaramuzza, "Perception-aware path planning," *arXiv preprint arXiv:1605.04151*, 2016.

[19] Z. Zhang and D. Scaramuzza, "Perception-aware receding horizon navigation for mavs." IEEE, 2018, pp. 2534–2541.

[20] ——, "Beyond point clouds: Fisher information field for active visual localization." IEEE, 2019, pp. 5986–5992.

[21] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," 2015, pp. 2938–2946.

[22] D. S. Chaplot, E. Parisotto, and R. Salakhutdinov, "Active neural localization," *arXiv preprint arXiv:1801.08214*, 2018.

[23] M. Lodel, B. Brito, A. Serra-Gómez, L. Ferranti, R. Babuska, and J. Alonso-Mora, "Where to look next: Learning viewpoint recommendations for informative trajectory planning." IEEE, 2022, pp. 4466–4472.

[24] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," 2018. [Online]. Available: https://arxiv.org/abs/1712.07629

[25] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021.

[26] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[27] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[28] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," 2018, pp. 8601–8610.

[29] I. A. Şucan, M. Moll, and L. E. Kavraki, "The Open Motion Planning Library," *IEEE Robotics & Automation Magazine*, vol. 19, no. 4, pp. 72–82, December 2012, https://ompl.kavrakilab.org.