

# Active Visual Localization for Multi-Agent Collaboration: A Data-Driven Approach

Matthew Hanlon\* Boyang Sun\* Marc Pollefeys Hermann Blum

**Abstract**—Rather than having each newly deployed robot create its own map of its surroundings, the growing availability of SLAM-enabled devices provides the option of simply localizing in a map of another robot or device. In cases such as multi-robot or human-robot collaboration, localizing all agents in the same map is even necessary. However, localizing e.g. a ground robot in the map of a drone or head-mounted MR headset presents unique challenges due to viewpoint changes. This work investigates how active visual localization can be used to overcome such challenges of viewpoint changes. Specifically, we focus on the problem of selecting the optimal viewpoint at a given location. We compare existing approaches in the literature with additional proposed baselines and propose a novel data-driven approach. The result demonstrates the superior performance of our data-driven approach when compared to existing methods, both in controlled simulation experiments and real-world deployment.

## I. INTRODUCTION

Visual localization and mapping systems are by now ubiquitous around humans. They are part of every smartphone, to e.g. improve localization in GPS denied environments, they are used in VR and AR headsets, in cars, and of course in robots. In parallel to this rollout, more and more environments get mapped, and localization becomes an interesting cloud service where any agent can send its observation and gets it localized in a pre-existing map. This can even be achieved without violating privacy [1]. However, it raises the question how well devices and robots can be localized if the map was created from a different kind of device and possibly from quite different points of view, as illustrated in Figure 2.

As an example of this larger question, this paper delves into the specific scenario of localizing a new agent, such as a ground robot, in a pre-existing map. This application holds considerable practical value, as it obviates the need to re-map an entire building if a suitable map already exists. Furthermore, for seamless collaboration between human-robot teams, the ability to localize mobile devices and robots within a shared map becomes imperative, as highlighted in previous studies [2]–[4]. However, accurately localizing a robot in a map created with a head-mounted camera rig introduces its own challenges. These challenges stem from the use of diverse sensor devices and are compounded by significant variations in viewpoint between the mapping trajectory and the operational height of the robot. Such variations result in multiple causes that diminishes the localization

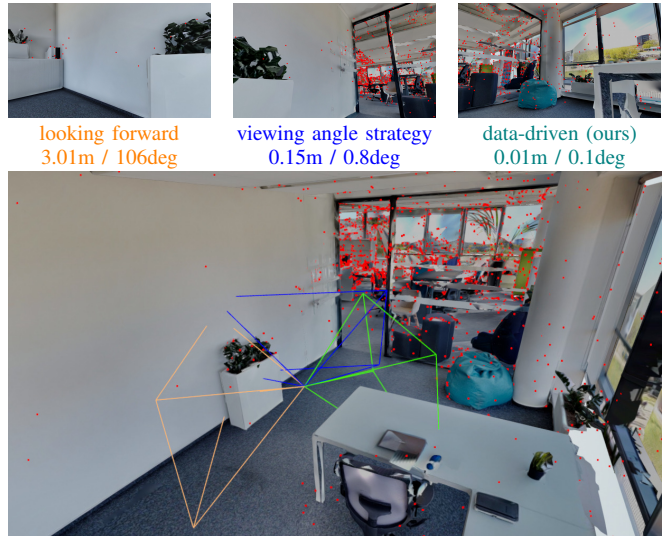


Fig. 1: **Viewpoint Selection** of three methods that run visual localization at the same location with respect to the built map (landmarks in red). The passive strategy of looking *forward*, and a strategy inspired by [5] to maximize the similarity with the *viewing angle* towards the landmark during mapping both result in higher localization error than our data-driven *viewpoint transformer (VPT)* approach.

performances, such as reduced visual overlap. This issue becomes particularly prominent in the case of ground robots, including quadrupeds, where obstacles such as chairs, tables, and furniture frequently obscure substantial portions of the robot’s environment.

Many works have studied how to better localize a given image within the map. However, in contrast to always trying to achieve the best from a given viewpoint, robots possess the valuable ability to autonomously select viewpoints. This leads us to investigate whether active viewpoint selection can effectively address the challenges associated with cross-agent visual localization. In the literature, this concept is widely recognized as active perception, and within our specific context, it is referred to as Active Visual Localization. The core objective of Active Visual Localization is to determine the camera pose within an existing map representation of the environment, in order to improve localization accuracy. A common approach involves assessing the localization utility of various viewpoints, typically through metrics like the Fisher Information Metric (FIM), or a combination of hand-crafted heuristics. While extensive work has been directed towards integrating these utility calculations into planning

\*equal contribution

All authors are with the Computer Vision and Geometry Lab at ETH Zürich.



Fig. 2: **Difference in perspective** between a head-mounted sensor rig used for mapping (left) and a ground robot (right) deployed for localization.

frameworks, a relatively less studied component is how the utility value itself can be better estimated, particularly in scenarios involving the unique challenges discussed earlier.

This work delves into the exploration of an effective utility function to actively enhance visual localization of a robot in a mapped environment. Our primary focus lies in evaluating established viewpoint selection and assessment criteria, while introducing a novel data-driven approach to viewpoint scoring. The key contributions of this paper are:

- **A novel data-driven approach** to viewpoint scoring resp. selection for active localization
- **Comparison and thorough evaluation** of viewpoint selection methods for the problem of visual localization between heterogeneous agents
- **Real-world validation** of our findings by integrating the viewpoint selection into an active viewpoint planner for a robot with an arm-mounted camera

## II. RELATED WORK

Active Vision describes the case where an agent has the ability to move visual sensors with the goal of improving the performance of perceptual tasks [6]. This concept finds application across various domains where visual information is utilized, such as scene exploration [7,8], data collection [9, 10], inspection [11], and active learning [12,13]. The case studied in this work focuses on the task of improving visual localization by choosing the most informative viewpoint at a given position, which can be considered a subset of Active Vision, named Active Visual Localization. Numerous existing works have studied how informative metrics can be incorporated into motion planning frameworks with the goal of optimizing robot motion to maximize localization accuracy [14]–[16]. Within the scope of this research, our focus is on the specific task of augmenting visual localization by selecting the most informative viewpoint at a given position.

Regarding viewpoint selection in the realm of active vision, early approaches primarily rely on handcrafted metrics to gauge the uncertainty or reliability of a given viewpoint in contributing to the system’s state [17]–[23]. Some methods take a different path by deriving evaluation metrics from the metric map and constructing a global utility map. Authors of [24] propose a solution rooted in Fisher information theory [25,26]. The central task they explore consists of determining the amount of information a viewpoint from a

given pose will contribute to the localization process. They develop a novel map representation that enables efficient computation of the Fisher information for 6-DoF visual localization, known as the Fisher Information Field. Similar ideas of using Fisher Information for viewpoint selection have also been explored in recent works [23,27,28]. However, it’s important to note that these metrics often rely on heuristics, necessitate the design of specific handcrafted utility functions, and may have limitations in their representation capabilities for diverse and complex scenarios.

Another category of works introduces additional vision tasks to enhance the metric extraction process, notably incorporating semantic information [29]. For instance, the work of [30] aims to improve navigation performance by including semantic information, in order to discern perceptually informative areas of the environment. This kind of work enriches the semantic understanding capability, however highly relies on the performance of the semantic module, and usually requires prior knowledge to link certain semantic classes that clearly correlate to visual informativeness.

A separate line of research focus on using data-driven approaches to active visual localization. These works mostly choose to formulate the problem as a reinforcement learning task, tightly integrating metric determination with robot execution [31,32]. As an example, [33] trains an information-aware policy to find traversable paths as well as reduce the uncertainty of the environment. The learning-based model significantly enhance the robot’s comprehension of its surroundings. However, it’s worth noting that these models typically demand substantial computational resources for training and may require the simplification of the environment model to prevent over-fitting.

In this work, we try to combine the advantages of both the data-driven approach and the viewpoint scoring scheme by formulating viewpoint selection as a classification problem.

## III. METHOD

**Problem Statement:** Let  $\mathbf{x} = (\mathbf{p}, \mathbf{q})$ , where  $\mathbf{p} \in \mathcal{R}^3$  and  $\mathbf{q} \in so(3)$  are the position and orientation of the robot sensor. Given a map representation of the environment  $\mathcal{M}$  and a prior of the robot position  $\hat{\mathbf{p}}$ , our goal of the active visual localization is to find a viewpoint at that location, i.e., the orientation for the robot sensor, such that the visual localization method returns the most accurate estimation  $\bar{\mathbf{x}}$  at that location, i.e.,

$$\mathbf{q}^* = \arg \min_{\mathbf{q}} \|\bar{\mathbf{x}} - \mathbf{x}\| \quad (1)$$

$$= \arg \min_{\mathbf{q}} \|\text{loc}(\mathcal{M}, \mathcal{O}(\mathbf{x})) - \mathbf{x}\| \quad (2)$$

Here  $\text{loc}(\cdot)$  refers to different localization methods, which often take the observation  $\mathcal{O}$ , captured at the current pose  $\mathbf{x}$ , and localize it against the given map representation  $\mathcal{M}$ .

In this paper, we construct  $\mathcal{M}$  that combines the landmark point cloud  $\mathcal{M}_l$  and the Truncated Signed Distance Function (TSDF)  $\mathcal{M}_t$  of the environment. Any viewpoint selection policy  $\pi(\cdot)$  then takes  $\mathcal{M} = (\mathcal{M}_l, \mathcal{M}_t)$  as prior knowledge

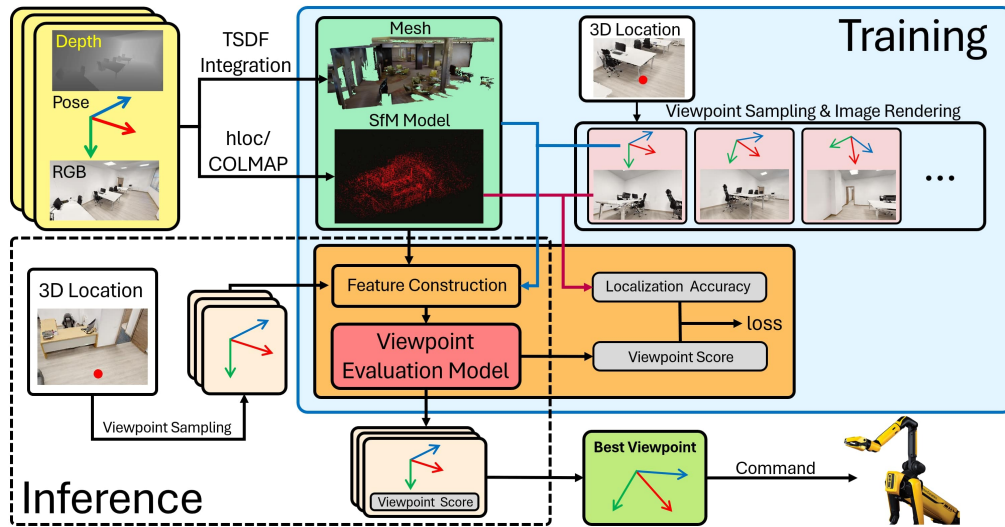


Fig. 3: **Overview of the proposed active localization approach** The core of our approach is the learning-based viewpoint evaluation model. This model processes input features derived from an established Structure-from-Motion model alongside a camera viewpoint. It predicts the likelihood of the given viewpoint being effective for visual localization. In practice, when deployed, multiple viewpoints are sampled and assessed at a particular 3D location. The viewpoint that receives the highest predicted score is then chosen as the optimal one to execute for the robot.

of the environment, and selects a viewpoint for a certain estimated position:

$$\mathbf{q}^* = \pi(\mathcal{M}, \hat{\mathbf{p}}) \quad (3)$$

As an example, one of the baseline methods we implement takes the idea from [24]:

$$\mathbf{q}^{\text{FIM}} = \pi_{\text{FIM}}(\mathcal{M}, \hat{\mathbf{p}}) = \arg \max_{\mathbf{q}} \sum_{i \in \mathcal{M}_t} v(\mathbf{x}, i) \mathbf{I}_i \quad (4)$$

where  $v(\cdot)$  is the binary visibility of landmark  $i$  and  $\mathbf{I}_i$  is the Fisher Information Metric (FIM) of observing landmark  $i$ . The visibility  $v(\cdot)$  is determined using TSDF map  $\mathcal{M}_t$ .

We supplement this with two additional simple baselines:

$$\pi_{\text{max}}(\mathcal{M}, \hat{\mathbf{p}}) = \arg \max_{\mathbf{q}} \sum_{i \in \mathcal{M}_t} v(\hat{\mathbf{x}}, i) \quad (5)$$

$$\pi_{\text{angle}}(\mathcal{M}, \hat{\mathbf{p}}) = \arg \max_{\mathbf{q}} \sum_{i \in \mathcal{M}_t} v_{\text{angle}}(\hat{\mathbf{x}}, i) \quad (6)$$

where  $\pi_{\text{max}}$  selects simply the view with the maximum visible landmarks, and  $\pi_{\text{angle}}$  uses a stricter visibility criterion inspired by [5] that only considers a landmark visible if its relative location to  $\mathbf{x}$  is similar to those poses from which it was seen during mapping.

We propose a data-driven approach, adhering to a "sample-and-evaluate" framework as depicted in Figure 3. Our process involves gathering simulated data to train our viewpoint evaluation model, which is then tested through both simulated scenarios and real-world experiments. In the inference phase, we sample multiple viewpoints at a given location. For each viewpoint, we assemble its input feature vector by utilizing the landmark point cloud derived from the Structure-from-Motion (SfM) model. The rest of this section will detail our approaches, with the focus on constructing the feature vectors and the viewpoint evaluation model.

**Data-driven viewpoint evaluation** For each viewpoint, we collect the following essential information:

- Its distance to every landmark of the map.
- The viewing angle between every landmark and the principle axis.
- The minimum and maximum distance and viewing angle per landmark has, with respect to the camera frame during the mapping stage
- Pixel coordinates of every landmark in the camera frustum of  $\mathbf{x}$
- The number of landmarks in the previously seen angle range
- Its corresponding DINO [34] appearance features for every landmark

To collect training data, we generate for every viewpoint the above information, as well as the ground truth pose and the result of the visual localization method at that viewpoint. Based on this data, we train our model to classify whether a viewpoint has been localized within an error threshold. In particular, we distinguish between two methods, one is based on Multi-Layer Perceptron (MLP)  $\pi_{MLP}$ , and the other one is based on the Transformer  $\pi_{VPT}$ . The way of encoding the collected information differs respectively. We design lightweight models in order to maintain online capability. Both kinds of models end with a softmax layer to determine the viewpoint score for classification.

The model based on MLP requires a fixed input dimensionality. However, the number of landmarks varies across different viewpoints, we set a feature aggregation step in the model of  $\pi_{MLP}$ . For each information in the list above, we build a histogram to aggregate the corresponding value of all the filtered landmarks. For the pixel coordinates, we aggregate the information in a 2D heatmap instead of a 1D histogram. DINO features cannot be processed in the MLP-

based model.

The constraints of fixed input dimensionality also encourages our motivation to investigate the transformer architecture. With the transformer, our inputs can be provided in a per-landmark fashion, which also allows for the inclusion of features that cannot be easily aggregated, such as DINO appearance per landmark.

Upon obtaining the trained model, for a specified point  $p$ , we assess all the sampled viewpoints. Each candidate is allocated a localization score, reflecting its efficacy in visual localization. Subsequently, we select the viewpoint that is awarded the highest score. For instance:

$$\pi_{\text{VPT}}(\mathcal{M}, \hat{p}) = \arg \max_q f_{\text{VPT}}(\mathbf{F}(\hat{x}, \mathcal{M})) \quad (7)$$

where  $\mathbf{F}$  are the input features, and  $f_{\text{VPT}}(\cdot)$  returns the output score from the viewpoint transformer (VPT) model.

#### IV. EXPERIMENTS

We develop a comprehensive pipeline for data collection aimed at training, validation, and testing across simulated and real-world settings, and we carry out a variety of experiments. Additionally, we implement several baseline methods for comparison. The data collection and model training is mostly done within the simulation, details can be found in IV-A and IV-B. We present all baseline methods alongside our evaluation strategy in IV-C. Finally, our real-world experiment findings are detailed in IV-D.

##### A. Data Generation

**Simulated Scenario** We focus on indoor scenarios, and choose a selection of scenes from the Habitat-Matterport 3D (HM3D) dataset [35]. HM3D provides high-quality 3D reconstructions of real-world indoor environments with textured 3D meshes. We selected nine scenes, as depicted in Figure 4, and imported them into NVIDIA Isaac Sim. Subsequently, we manually captured controlled trajectories using a simulated camera setup that mimics the Microsoft HoloLens 2 [36], ensuring the camera was positioned at a height akin to that of a human.

**Data Collection** To acquire  $\mathcal{M}_t$ , we utilize the depth images captured by sensors and conduct TSDF (Truncated Signed Distance Field) integration<sup>1</sup>. Both the generated meshes and the occupancy maps can be used for occlusion and collision checking.

To obtain  $\mathcal{M}_t$ , we use mapping and localization frameworks from hloc and COLMAP [37,38] to extract 2D local features from images and build a 3D landmark point cloud. The collected images from the simulated HoloLens 2 are fed into the pipeline to build the 3D landmarks map. The per-landmark feature vector  $\mathbf{F}$  are also created and attached to each landmark at this stage.

For running visual localization, We mainly use the localization module from hloc as the framework for  $loc(\cdot)$  in 2.

<sup>1</sup>Isaac Sim offers a direct method to ascertain the occupancy details of a 3D scene through its built-in functionality, which we also employ as an alternative approach.



Fig. 4: **Overview of the dataset.** The number in brackets following the designation corresponds to the index in the Habitat-Matterport 3D dataset. A small collection of scenes lead to great generalization capability of our model, thanks to our effective data point sampling method.

To recover the scale of the map, one option would be to create the reconstruction with known camera poses. However, this would not be representative of how this process could be done using real hardware, where poses would be estimated, for example, from a visual SLAM module. In our experiment, all maps are created without the exact camera poses. Instead, we build the reconstructions from the images alone and then using RANSAC to estimate the scale with respect to the ground-truth poses.

To collect training data point in the simulator, we create random camera paths at a height characteristic of robots, using a single RGB-D camera to mimic an onboard robot sensor. At each waypoint, we capture a set of viewpoint samples and store the images along with their exact viewpoint poses as ground truth. As illustrated in 1 and 2, we then employ the localization module for each sample, calculating the discrepancy between the estimated pose and the actual ground truth pose.

##### B. Model Training

We train both two kinds of models on the classification task of predicting whether a viewpoint will result in localization errors smaller than 0.1m and 1 deg. The primary preprocessing actions include normalizing all input features  $\mathbf{F}$  to a range between 0 and 1 and ensuring the dataset has a balanced distribution of positive and negative examples. The division of the training and validation datasets is illustrated in Figure 4. For every scene, we randomly generate 100 waypoints. At each waypoint, we sample 50 viewpoints, allowing complete rotation around the yaw axis while keeping the pitch angle within a degree range between -10 to 45. This procedure generates a training and validation dataset comprising 25,000 images and a test set containing 20,000 images.

##### C. Evaluation Strategy

As detailed in III, we implement three baseline strategies  $\pi_{\text{FIM}}$ ,  $\pi_{\text{max}}$ ,  $\pi_{\text{angle}}$ , and augment them with two additional simplistic approaches:  $\pi_{\text{forward}}$ , which directs the camera towards the next waypoint in the trajectory, and  $\pi_{\text{random}}$ , which chooses viewpoints at random. Following the methodology established during the training phase, random waypoints and viewpoint samples are generated for each evaluation

distance [m]		0.025	0.05	0.075	0.1	0.25	1.0	distance [m]		0.025	0.05	0.075	0.1	0.25	1.0
orientation [deg]		1	1	1	1	2	5	orientation [deg]		1	1	1	1	2	5
w/o occl. filt.	Forwards	62.57	77.25	80.84	81.99	85.23	86.68	w/o occl. filt.	Forwards	37.97	55.96	62.62	65.71	74.70	79.22
	Random	54.84	69.31	72.60	73.90	78.29	80.29		Random	31.21	48.66	55.57	57.80	65.46	70.23
	max	60.43	76.05	79.29	80.64	82.93	84.98		max	33.05	50.94	57.60	60.19	67.99	73.01
	angle	74.35	85.58	87.28	87.72	89.07	90.27		angle	41.20	55.86	60.59	63.07	70.73	75.60
	MLP	76.25	86.78	88.37	88.82	90.42	91.52		MLP	<b>51.09</b>	67.84	73.56	75.84	81.61	84.69
	VPT	74.75	87.67	90.67	91.27	92.96	93.86		VPT	44.23	62.77	69.48	72.12	78.98	82.85
	VPT + DINO	69.16	84.68	87.97	88.82	90.92	92.07		VPT + DINO	39.02	58.45	65.16	68.34	76.44	80.27
w/ occl. filt.	max	70.21	84.98	88.02	89.42	91.37	92.22	w/ occl. filt.	max	46.42	67.84	74.20	76.54	84.05	<b>88.72</b>
	angle	78.34	88.22	90.02	90.37	91.52	92.56		angle	50.80	<b>70.78</b>	75.75	78.03	84.59	88.52
	FIM	65.17	78.54	81.79	83.53	86.83	88.67		FIM	43.34	63.02	69.53	70.87	78.43	82.85
	MLP	<b>79.69</b>	89.72	91.42	92.17	93.66	94.91		MLP	<b>51.49</b>	<b>70.43</b>	<b>76.94</b>	<b>79.47</b>	<b>84.94</b>	88.02
	VPT	<b>79.54</b>	<b>90.97</b>	<b>93.31</b>	<b>93.76</b>	<b>94.61</b>	<b>95.21</b>		VPT	41.70	64.12	71.67	74.75	81.76	85.74
	VPT + DINO	78.64	<b>90.52</b>	<b>92.56</b>	<b>92.91</b>	<b>94.56</b>	<b>95.36</b>		VPT + DINO	47.27	69.48	<b>76.49</b>	<b>79.37</b>	<b>85.14</b>	<b>89.26</b>
	Best Possible	96.31	97.46	97.70	97.70	98.05	98.10		Best Possible	88.52	92.84	93.74	94.28	97.02	97.66

TABLE I: **Evaluation of viewpoint strategies** in the simulated testing environments using both **SuperPoint**(Left) and **SIFT**(Right) features. Recall percentages are shown at varying distance and orientation thresholds, highlighting the **best** and **second-best** performing methods. Rows marked as ‘with occlusion filter’ are those where landmarks are filtered with the occlusion. The bottom row is an oracle method that selects the viewpoint with smallest possible error among all samples.

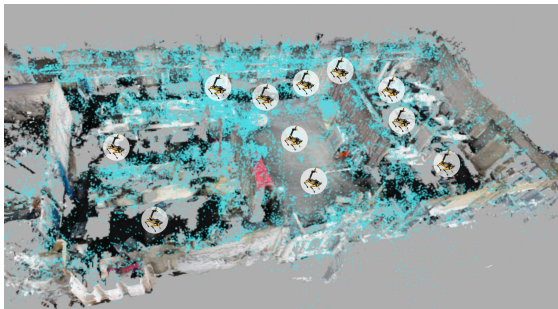


Fig. 5: **The constructed map for real-world evaluation** The landmarks point cloud (blue) is aligned with the environment mesh. Evaluated locations are shown as mini robots.

setting. Each sampled viewpoint is evaluated according to the different viewpoint-selection strategies, recording the localization error for the selected viewpoints. To establish an upper limit for the attainable localization precision at a given position, the minimal error observed from the localization for each sample is also recorded and treated as the pseudo-optimal viewpoint selection strategy.

#### D. Real-world Deployment and Test

**Setup** To further validate the results obtained from the simulation experiments, the best-performing methods are implemented into a ROS-compatible planning module and deployed on a quadruped robot with a robotic arm that contains a calibrated color camera in its end effector, allowing different viewpoints to be viewed for a given body position. We use a HoloLens 2 to build a map of an indoor environment, as shown in Figure 5.

**Data Collection** In contrast to the simulated environment where ground-truth poses are readily available, real-world deployments lack these precise positional references. To assess the effectiveness of various viewpoint-choosing strategies in such settings, two kinds of positional data are essential: Each viewpoint planner must first obtain an initial estimate  $\hat{p}$ , which is utilized in 3 to determine an appropriate viewing direction at the current location. Following the

selection of a viewing direction, the robot adjusts its arm to capture an image, which is then localized in relation to the pre-existing map. Subsequently, a ground-truth pose is necessary to validate the accuracy of this localization

To measure accurate poses in the map created from the HoloLens recording, we use a combination of visual localization and AprilTag fiducial markers [39]. Once a map of the environment has been created, an AprilTag is affixed to a static location. We capture several high-quality images of the marker, including its surroundings, using a calibrated DSLR camera. The AprilTag marker can be accurately localized within the captured images, and by ensuring that the images contain enough of the surroundings, they can also be localized in the map using visual localization. This two-step process provides an estimated location of the AprilTag marker within the map for each image. An average of the estimated locations is taken in order to minimize the effect of errors in the tag or localization. This process gives us accurate poses of fiducial markers with respect to the captured landmark map.

For any waypoint on which we want to evaluate the viewpoint strategies, we initialize the robot at the location where we placed a fiducial marker and walk the robot to the investigated waypoint. We estimate  $\hat{p}$  of that waypoint from observing the fiducial marker with the robot camera and integrating the odometry to the investigated waypoint. This resembles a realistic odometry-based localization prior that can be fed into the viewpoint-choosing strategies.

## V. RESULTS

We evaluate our approaches and the baseline methods in both our simulation pipeline and the real world. Table I shows the quantitative results from the simulation. To demonstrate the generalization ability of our approaches across different feature descriptors, we evaluate all methods using both SuperPoint feature [40] and SIFT feature [41], whereas both of our scoring models are only trained with the SuperPoint feature. The result in the table shows that,

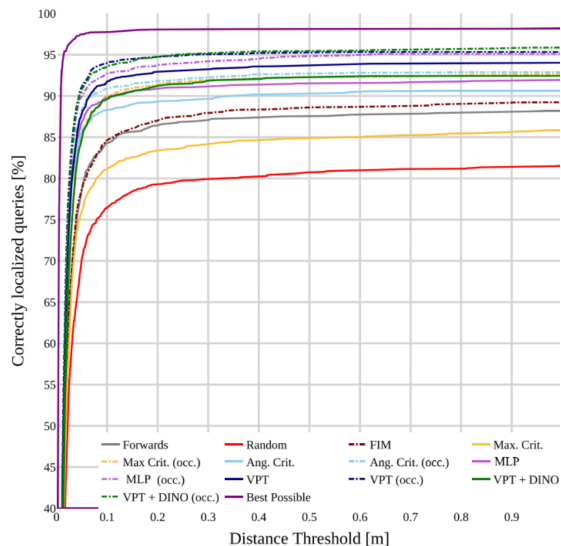


Fig. 6: **Cumulative distribution of position errors** of the evaluation points in the testing environments using **SuperPoint** feature. Methods or versions of methods that rely on the environment mesh in order to determine landmark occlusion are plotted using  $- \cdot - \cdot -$  lines.

with SuperPoint feature, our methods perform the best under every error level, even though our models are only trained to classify the 0.1 m, 1 deg error level. When we switch from SuperPoint to SIFT, recall percentages drop for all the methods, however, our data-driven methods still have the highest recall at almost every error level.

To explore the impact of DINO features and occlusion handling on model performance, we conduct experiments with various model configurations, including MLP-based and VPT-based models that do not pre-filter based on occlusion information, as well as VPT-based models that do not incorporate DINO features in their input. The findings reveal that neglecting occlusion leads to diminished performance for both model types, a trend that is also observed in baseline methods. Interestingly, the VPT-based model demonstrates the ability to exceed the efficacy of traditional non-data-driven approaches at its designated training threshold, even without occlusion considerations. We accumulate the correctly-localized waypoints along with different distance thresholds and show the result in Figure 6, where it is easier to verify that our learning-based approaches outperform the other baseline methods with both SuperPoint and SIFT features.

Results from the 12 waypoints (see Figure 5) in our real-world experiment are shown in Table II. Although the sample size is relatively small, we still see the dominating performance of the VPT-based model, which shows its generalization ability to the real world. Besides, our final approach shows great online performance. It achieves to successfully select viewpoints from 100 candidates in less than one second, running on a regular workstation with a NVIDIA GeForce RTX 2080 GPU.

The evaluation results highlight the effectiveness of data-driven methods in comparison to hand-crafted information

distance [m]	0.25	0.5	1.0
orientation [deg]	2.0	3.5	5.0
Arm stowed	0.0	25.0	41.67
ForwardS	16.67	50.0	58.33
Robot odometry	0.0	33.33	66.67
Random	8.33	41.67	58.33
FIM	8.33	50.0	75.0
max	8.33	66.67	83.33
angle	8.33	58.33	66.67
MLP	0.0	58.33	58.33
VPT	33.33	83.33	91.67

TABLE II: **Evaluation of viewpoint strategies** in the real-world environment with the quadruped robot. The table layout is the same as Table I

metrics. Learning decision boundaries for landmark features and encoding diverse information enable data-driven approaches to outperform traditional heuristic metrics. Despite being trained with a limited dataset, these scoring models exhibit robustness across various scenarios, different local features, and both simulated and real-world environments. This underscores the potential of data-driven approaches, which adopt a less heuristic approach.

We note variations in performance between the two machine learning methodologies. The transformer model exhibits reduced generalization across distinct feature descriptors yet demonstrates superior transferability from simulated data to real-world applications. Conversely, the MLP-based model and the VLP-based model employing DINO underperform with SuperPoint features and in real-world evaluations but show enhanced adaptability to SIFT features. This suggests that in these scenarios, they benefit from a more substantial semantic prior provided by DINO, which in turn offers greater generality across various feature types.

In conclusion, our proposed data-driven light transformer-based model exhibits optimal performance when evaluated on the same feature descriptors as those used during training. The inclusion of occlusion filtering based on mesh reconstruction enhances the performance of all approaches. It’s important to note that in cases where geometric data isn’t available, the proposed data-driven approach still produces superior results. This could be attributed to the learning process capturing information about the angle ranges from which landmarks were observed, which indirectly encodes geometric information.

## VI. CONCLUSION

This work addresses the challenge of localizing ground robots within an existing map constructed from devices with varying perspectives. We introduce a novel data-driven approach to explore effective utility functions for this task and evaluate it alongside diverse viewpoint selection methods in the literature. Experiment results show that our approach greatly improves robot localization ability within a known point-cloud map in the presence of large viewpoint changes and occlusion from ground-level obstacles. These improvements are observed in both simulated and real-world environments.

## REFERENCES

- [1] P. Speciale, J. L. Schonberger, S. B. Kang, S. N. Sinha, and M. Pollefeys, "Privacy preserving image-based localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5493–5503.
- [2] J. Chen, B. Sun, M. Pollefeys, and H. Blum, "A 3d mixed reality interface for human-robot teaming," in *submission*, 2023.
- [3] O. Erat, W. A. Isop, D. Kalkofen, and D. Schmalstieg, "Drone-augmented human vision: Exocentric control for drones exploring hidden areas," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1437–1446, 2018.
- [4] C. Reardon, K. Lee, and J. Fink, "Come see this! augmented reality to enable human-robot cooperative search," in *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2018, pp. 1–7.
- [5] A. J. Davison and D. W. Murray, "Simultaneous localization and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865–880, 7 2002.
- [6] J. Aloimonos, "I. weiss and a. bandyopadhyay: Active vision," in *Proc. 1st Int. Joint Conf. Computer Vision*, 1987, pp. 35–54.
- [7] L. Schmid, M. Pantic, R. Khanna, L. Ott, R. Siegwart, and J. Nieto, "An efficient sampling-based method for online informative path planning in unknown environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1500–1507, 2020.
- [8] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," in *International Conference on Learning Representations (ICLR)*, 2020.
- [9] J. Rückin, F. Magistri, C. Stachniss, and M. Popović, "An informative path planning framework for active learning in uav-based semantic mapping," *arXiv preprint arXiv:2302.03347*, 2023.
- [10] K. Ye, S. Dong, Q. Fan, H. Wang, L. Yi, F. Xia, J. Wang, and B. Chen, "Multi-robot active mapping via neural bipartite graph matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 839–14 848.
- [11] J. Xing, G. Cioffi, J. Hidalgo-Carrió, and D. Scaramuzza, "Autonomous power line inspection with drones via perception-aware mpc," in *IEEE/RSJ International Conference on Intelligent Robots (IROS)*, October 2023.
- [12] R. Zurbrügg, H. Blum, C. Cadena, R. Siegwart, and L. Schmid, "Embodied active domain adaptation for semantic segmentation via informative path planning," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8691–8698, 2022.
- [13] J. Rückin, F. Magistri, C. Stachniss, and M. Popović, "Semi-supervised active learning for semantic segmentation in unknown environments using informative path planning," *IEEE Robotics and Automation Letters*, 2024.
- [14] N. Roy, W. Burgard, D. Fox, and S. Thrun, "Coastal navigation-mobile robot navigation with uncertainty in dynamic environments," in *Proceedings 1999 IEEE international conference on robotics and automation (Cat. No. 99CH36288C)*, vol. 1. IEEE, 1999, pp. 35–40.
- [15] G. Costante, C. Forster, J. Delmerico, P. Valigi, and D. Scaramuzza, "Perception-aware path planning," *arXiv preprint arXiv:1605.04151*, 2016.
- [16] Z. Zhang and D. Scaramuzza, "Perception-aware receding horizon navigation for mavs," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2534–2541.
- [17] S. A. Sadat, K. Chutskoff, D. Jungic, J. Wawerla, and R. Vaughan, "Feature-rich path planning for robust navigation of MAVs with MonoSLAM," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3870–3875, 9 2014.
- [18] A. Yamashita, K. Fujita, T. Kaneko, and H. Asama, "Path and viewpoint planning of mobile robots with multiple observation strategies," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 4, 2004, pp. 3195–3200 vol.4.
- [19] A. Kim and R. M. Eustice, "Perception-driven navigation: Active visual slam for robotic area coverage," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 3196–3203.
- [20] D. Fontanelli, P. Salaris, F. A. Belo, and A. Bicchi, "Visual appearance mapping for optimal vision based servoing," in *Experimental Robotics: The Eleventh International Symposium*. Springer, 2009, pp. 353–362.
- [21] Z. Zhang and D. Scaramuzza, "Perception-aware receding horizon navigation for mavs," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2534–2541.
- [22] C. Papachristos, S. Khattak, and K. Alexis, "Uncertainty-aware receding horizon exploration and mapping using aerial robots," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4568–4575.
- [23] J. Lim, N. Lawrance, F. Achermann, T. Stastny, R. Bähnemann, and R. Siegwart, "Fisher information based active planning for aerial photogrammetry," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 1249–1255.
- [24] Z. Zhang and D. Scaramuzza, "Fisher information field: an efficient and differentiable map for perception-aware planning," *arXiv preprint arXiv:2008.03324*, 2020.
- [25] H. J. S. Feder, J. J. Leonard, and C. M. Smith, "Adaptive mobile robot navigation and mapping," *The International Journal of Robotics Research*, vol. 18, no. 7, pp. 650–668, 1999.
- [26] A. A. Makarenko, S. B. Williams, F. Bourgault, and H. F. Durrant-Whyte, "An experiment in integrated exploration," in *IEEE/RSJ international conference on intelligent robots and systems*, vol. 1. IEEE, 2002, pp. 534–539.
- [27] A. Kim and R. M. Eustice, "Active visual slam for robotic area coverage: Theory and experiment," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 457–475, 2015.
- [28] I. Abraham and T. D. Murphey, "Active learning of dynamics for data-driven control using koopman operators," *IEEE Transactions on Robotics*, vol. 35, no. 5, pp. 1071–1083, 2019.
- [29] Y. Tao, X. Liu, I. Spasojevic, S. Agarwa, and V. Kumar, "3d active metric-semantic slam," 2023.
- [30] L. Bartolomei, L. Teixeira, and M. Chli, "Semantic-aware Active Perception for UAVs using Deep Reinforcement Learning," *IEEE International Conference on Intelligent Robots and Systems*, pp. 3101–3108, 2021.
- [31] D. S. Chaplot, E. Parisotto, and R. Salakhutdinov, "Active neural localization," *arXiv preprint arXiv:1801.08214*, 2018.
- [32] Q. Fang, Y. Yin, Q. Fan, F. Xia, S. Dong, S. Wang, J. Wang, L. J. Guibas, and B. Chen, "Towards accurate active camera localization," in *European Conference on Computer Vision*. Springer, 2022, pp. 122–139.
- [33] M. Lodel, B. Brito, A. Serra-Gómez, L. Ferranti, R. Babuška, and J. Alonso-Mora, "Where to look next: Learning viewpoint recommendations for informative trajectory planning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4466–4472.
- [34] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
- [35] 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.
- [36] P. Hübner, K. Clintworth, Q. Liu, M. Weinmann, and S. Wursthorn, "Evaluation of hololens tracking and depth sensing for indoor mapping applications," *Sensors*, vol. 20, no. 4, p. 1021, 2020.
- [37] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.
- [38] J. L. Schonberger and J. M. Frahm, "Structure-from-Motion Revisited," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 4104–4113, 12 2016.
- [39] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 3400–3407.
- [40] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
- [41] D. G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, 1999.