IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

VISTA: Monocular Segmentation-Based Mapping for Appearance and View-Invariant Global Localization

Hannah Shafferman[1,2,3], Annika Thomas[1], Jouko Kinnari[4], Michael Ricard[3], Jose Nino[3], and Jonathan P. How[1]

***Abstract*— Global localization is critical for autonomous navigation, particularly in scenarios where an agent must localize within a map generated in a different session or by another agent, as agents often have no prior knowledge about the correlation between reference frames. However, this task remains challenging in unstructured environments due to appearance changes induced by viewpoint variation, seasonal changes, occlusions, and perceptual aliasing in homogeneous environments — known failure modes for traditional place recognition methods. To address these challenges, we propose VISTA (View-Invariant Segmentation-Based Tracking for Frame Alignment), a novel open-set, monocular global localization framework that combines: 1) a front-end, object-based, segmentation and tracking pipeline, followed by 2) a submap correspondence search, which exploits geometric consistencies between environment maps to align vehicle reference frames. VISTA enables consistent localization across diverse camera viewpoints and seasonal changes, without requiring any domain-specific training or finetuning. We evaluate VISTA on seasonal and oblique-angle aerial datasets, achieving up to a 69% improvement in recall over baseline methods. Furthermore, we maintain a compact object-based map that is only 0.6% the size of the most memory-conservative baseline, making our approach capable of real-time implementation on resource-constrained platforms.**

I. Introduction

Autonomous vehicles operating in GNSS-denied environments—such as urban canyons, forests, indoor facilities, or adversarial settings—require robust localization [1]. Vision-based localization is particularly challenging due to large appearance variations caused by changes in viewpoint, season, lighting, shadows, spatial aliasing, and occlusions. In multi-agent settings, global localization—the task of estimating pose within a prebuilt map without an initial guess [2]—is crucial for aligning reference frames for coordinated perception and planning. However, differing viewpoints across agents often lead to severe appearance changes that challenge even state-of-the-art methods [3].

Despite its prevalence, the off-nadir (oblique) camera viewpoint remains understudied. This viewpoint is common in UAV teams, where sensor orientation varies across platforms, e.g., forward-facing monocular cameras on small UAVs used for obstacle avoidance and localization. Oblique viewpoints introduce object distortion, degrade feature matching, and amplify odometry uncertainty, creating substantial challenges for accurate localization.

*This work is supported by the Charles Stark Draper Laboratory, Inc. and funded in part by ONR.

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA.{hshaff, annikat, jhow}@mit.edu

[2]Draper Scholar.

[3]Charles Stark Draper Laboratory, Inc., Cambridge, MA, USA.

[4]Work done while at Saab Finland Oy, now at NestAI Oy {jouko.kinnari@nestai.com}

Fig. 1. VISTA segments and tracks objects through a video stream from different monocular camera orientations without domain-specific fine tuning in a zero-shot framework to build sparse environment maps. We utilize a data association framework using a geometric submap implementation, which leverages the underlying environment geometry to identify object correspondences between maps. Example segmented imagery from Highbay dataset.

Visual appearance-based methods [4], [5] detect and match local or global features, but these features vary significantly across seasons, lighting, and viewpoints [6], [7]. While recent methods [3], [8]–[11] improve robustness to appearance changes, they primarily target nadir or ground-view perspectives. This gap motivates methods that generalize to oblique viewpoints without requiring domain-specific training.

Onboard compute constraints further limit real-time localization solutions. Dense feature-based maps scale poorly in cluttered environments such as warehouse or forest environments. Object-based maps, though compact, require expensive global data association. This motivates our geometric submap strategy, which bounds correspondence search while preserving geometric consistency.

We propose VISTA (View-Invariant Segmentation-Based Tracking for Frame Alignment), a monocular global localization pipeline designed for severe appearance changes induced by seasonal variation and camera viewpoint shifts. VISTA uses open-set instance segmentation [12], [13] to build sparse, viewpoint-invariant 3D object maps (Fig. 1).

Prior object-based localization frameworks address complementary but limited aspects of this problem. ROMAN [11] requires RGB-D sensing, limiting use in monocular-only settings. SOS-Match [3] operates on monocular RGB and introduced open-set object segmentation but is restricted to nadir viewpoints and does not account for distortions or occlusions inherent to oblique viewpoints.

VISTA bridges these gaps by extending segmentation-based localization to off-nadir viewpoints, operating purely on monocular RGB, and introducing an auto-segmentation

tracker that leverages full object geometry for temporally consistent tracking. Combined with a geometric submap matcher, VISTA enables robust, scalable global localization across diverse conditions. These robustness properties are quantitatively demonstrated in Section IV through experiments on occluded oblique-viewpoint scenarios and perceptually-aliased seasonal datasets.

In summary, the contributions of this work include:

- A monocular auto-segmentation object tracker that maps objects across varying viewpoints, producing sparse, uncertainty-aware 3D representations for efficient and reliable global localization.
- A geometric submap correspondence search framework that incorporates object uncertainties and graph-theoretic association, yielding at least a $62\times$ reduction in computation time over most traditional and learning-based baselines, while maintaining significantly smaller maps (0.03%–0.6% of baseline map sizes). Full analysis and explanation are given in Section IV-F.
- Extensive evaluation on seasonal and oblique-viewpoint datasets, achieving up to 69% maximum recall improvement over traditional and learning-based visual place recognition (VPR) baselines.

II. Related Work

In this section, we review prior work on global localization in unstructured environments and highlight current limitations that motivate our approach for achieving robust performance under extreme environment appearance variation.

A. Map Representations

Successful global localization is highly dependent on a descriptive environment representation. Dense map representations (i.e. surfel maps [14], 3D Gaussian splats [15], NeRFs [16], or voxel maps [17]) encode complex geometric information about the environment but are computationally expensive and are not robust to domain shifts. Sparse maps, on the other hand, are lightweight and scalable environment reconstructions, which represent a scene as a set of 3D points [18]. Sparse maps are commonly either feature-based or object-based representations [19], [20]. Feature-based methods store low level distinct object features (i.e. corners, edges). Traditional feature detectors, such as SIFT [21], SURF [22], and ORB [23], are widely used to generate unique feature descriptors, but they rely on the assumption during matching that objects appear similarly between frames. Deep learning-based methods [7], [24]–[27] have been trained to detect, describe, and match features between frames to improve robustness to appearance changes caused by illumination or motion blur. However, their performance degrades under more extreme appearance changes and when the contents are perceptually similar between scenes [3], [5]. Our work builds upon prior works [3], [11], which leverage object-based environment representations. These object-based maps are lightweight and less sensitive to viewpoint variation and appearance or illumination changes, making them useful for information sharing in multi-agent systems.

B. Segmentation

Instance segmentation enables object-based mapping by providing semantics and precise object geometries within unstructured scenes [28]. The Segment Anything Model (SAM) [12] provides zero-shot segmentation for any object in an image. Segment Anything 2 (SAM2) [13] extends this capability to video segmentation by maintaining a memory of segment instances throughout the image sequence, providing robustness even in the presence of occlusions. Our approach builds upon these models by leveraging segment masks as descriptors for robust object-based tracking and re-identification throughout challenging scenes such as in the presence of viewpoint variations, seasonal changes, spatial aliasing, and occlusions.

C. Unstructured Environments

Many global localization works rely on distinct, repeatable features of urban settings (i.e. roads, street signs, lane markings) to simplify correspondence search for VPR [5], [29]–[31]. In contrast, unstructured environments lack such features and present challenging scenarios due to their irregular nature, repetitive terrain, lighting variation, and changing environmental conditions [32]–[34]. These features of unstructured environments make perception difficult and introduce challenges such as perceptual aliasing and occlusions. Open-set operation refers to a system's ability to handle inputs not encountered during training. Unstructured environments typically require such systems to handle many unknowns robustly. Our global localization pipeline leverages object-based sparse maps, which generalize well to unstructured environments compared to feature-based methods and offer robustness to visual appearance changes caused by varying environmental conditions.

III. Method

This section presents an overview of VISTA by first introducing our auto-segmentation object tracking and mapping pipeline followed by a discussion of our submap correspondence search frame alignment step. A visualization of the data flow is depicted in Fig. 2.

A. Notation

We use indices i, l, k, t, j for vehicles, image groups, detected objects, tracks, and detections respectively; scalars $N_l, N_{\text{obj}}, D_t, N_{\text{max}}, N_{\text{cand}}$ denote object counts per group, total tracked objects, detections per track, maximum submap objects, and candidate correspondences.

B. Auto-segmentation Object Tracking

Our object tracking pipeline relies solely on an image sequence as input, in contrast to previous work [3], which required camera pose estimates informed by IMU data for fundamental matrix filtering during segment tracking. Our tracker generates a set of object tracks, each consisting of a series of segment masks for every frame in which the object is detected.

We consider a set of vehicles indexed by i, each capturing an image sequence of length N_i frames. Since the video segmentation model [13] requires full sequence access, we enable real-time operation by processing images in batches of size $g_i \leq N_i$, tunable based on memory and computational constraints. Each image group is indexed by $l \in \{0, \ldots, \text{ceil}(N_i/g_i) - 1\}$.

For the first image in each group, we utilize a pre-trained image segmentation model [12] to generate object masks by

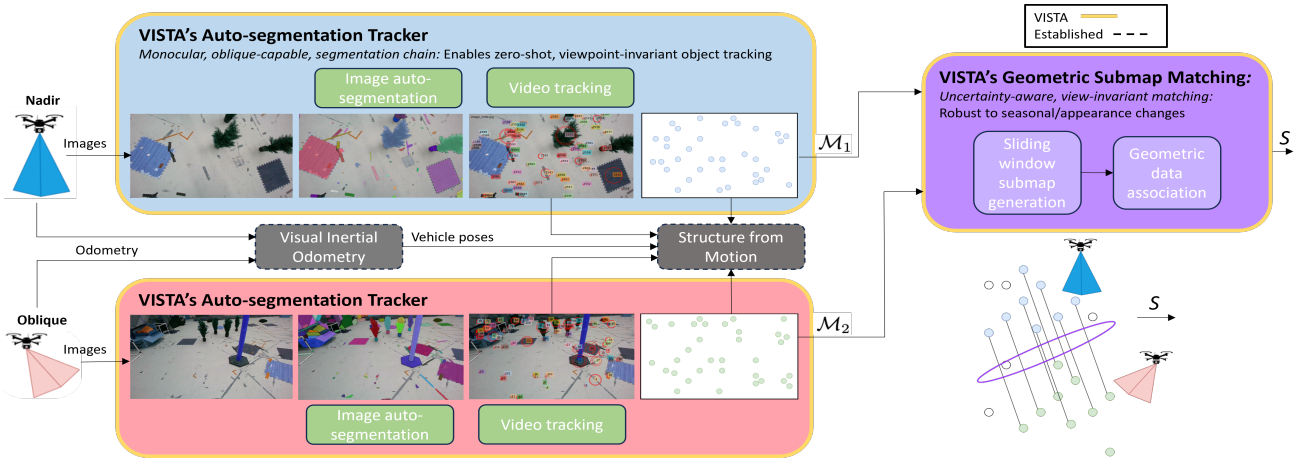

Fig. 2. System overview of VISTA's global localization pipeline. Gold-bordered regions denote novel contributions: (1) the auto-segmentation tracker employs a segmentation model inference chain to enable monocular, oblique-capable object tracking with zero-shot generalization and temporal consistency, and (2) the geometric submap matching module performs uncertainty-aware, view-invariant correspondence search robust to seasonal and appearance changes. Visual Inertial Odometry (VIO) and Structure from Motion (dashed gray boxes) are established components that provide camera poses and 3D object position estimates. Each vehicle builds a sparse environment map $\mathcal{M}_i$ from its trajectory, which is partitioned into overlapping submaps for efficient geometric data association between agents.

prompting the model with a 32×32 grid of points. This grid provides a dense, uniform sampling over each image while maintaining computational efficiency. We denote the number of detected objects in group l as N_l. The model returns a set of binary segment masks, $b_{k,l}$, where $k \in \{1, \ldots, N_l\}$, describing the spatial extent of each detected object.

These masks initialize the video segmentation model [13], which requires initial object definitions (points, boxes, or masks) to begin tracking. This chain of model inference enables a fully autonomous "auto-segmentation" pipeline where SAM's [12] zero-shot generalization provides accurate initial boundaries in unstructured or previously unseen environments, while SAM2 [13] maintains temporal consistency and occlusion robustness. By leveraging segment masks directly, we address the granularity issue [35], [36] faced by other pipelines, ensuring meaningful object-level video segmentation.

To account for objects entering and leaving the camera's field of view, we periodically reapply the image segmentation model [12] when the segmented area of the input image falls below a threshold $\theta_a = 0.5$. New objects are added to the tracked set. Existing tracked objects are not reassigned; instead, a spatial filtering step retains only new segments that significantly overlap with previously unsegmented regions. Redundant objects arising during transitions between image groups are resolved later during the correspondence search and do not affect tracking integrity.

For the N_l objects detected in image group l, the tracker outputs a set of binary segment masks $\{m_j\}$, where $j \in \{1, \ldots, N_l\}$. For each mask m_j, we extract the centroid θ_j as the 2D detection. We store these detections in a set $\mathcal{D}$, where each entry $\mathcal{D}[t]$ corresponds to a unique object track $t \in \{1, \ldots, N_{\text{obj}}\}$, with N_{obj} the total number of tracked objects. Similarly, $\mathcal{F}[t]$ stores the frames in which object t was detected.

We assume a generic Visual Inertial Odometry (VIO) system provides camera poses $T(t) \in SE(3)$. For reproducibility, our implementation uses ground-truth poses to simulate VIO estimates and is robust to typical pose errors. We simulated pose errors up to 0.3m translation and 3° rotation noise and found our triangulation and correspondence search remained robust. Using a Structure-from-Motion (SfM) triangulation approach, we reconstruct the 3D environment, estimating object positions and associated uncertainties using only RGB imagery.

We construct a factor graph $\mathcal{G}_t$ [37] for each object track t with at least $D_{\min} = 3$ detections, where nodes represent camera poses or object states and edges encode projection factors parameterized by detections and known camera calibration. The graph solves for 3D position by minimizing reprojection error. Our auto-segmentation tracker addresses the primary SfM challenge of data association and initialization [38] by providing robust tracking and triangulation-based initial estimates. Tracks where bundle adjustment fails are discarded as dynamic. The environment map $\mathcal{M}_i$ for each vehicle i contains estimated 3D positions μ_t and covariances for all tracked objects.

The uncertainty of each object's estimated 3D position is quantified using the marginal covariance computed from the factor graph's optimized posterior. After solving the nonlinear least squares problem via Levenberg–Marquardt optimization in GTSAM [38], we obtain the covariance for each landmark $\mathcal{L}_t$ from the inverse of the system's Hessian, marginalizing over all other variables. This marginal covariance $\Sigma_t = \text{Cov}(\mathcal{L}_\sqcup)$ reflects both measurement noise and geometric uncertainty from the observing camera poses. The mean and covariance (μ_t, Σ_t) are then stored in the environment map $\mathcal{M}_i$ for subsequent correspondence search, allowing uncertainty-aware matching across submaps.

Unlike prior work [3], VISTA does not maintain explicit size attributes for tracked objects. We found that purely geometric position-based matching is more robust to oblique viewpoint variations, where projected object sizes vary significantly with camera angle and distance. Additionally, this design choice contributes to VISTA's lightweight map representation by storing only position and uncertainty in-

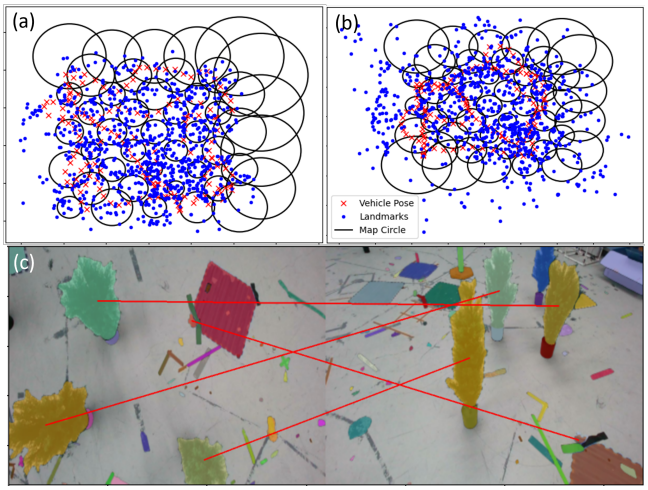

Fig. 3. Visualization of VISTA's object-based mapping and correspondence search. (a,b) Bird's-eye views of 3D object-based maps showing estimated object positions (blue dots), vehicle poses (red crosses), and overlapping sliding window submaps (black circles). (c) Correspondence search result from Highbay Experiment 1 showing geometrically consistent object matches (red lines) between nadir and oblique viewpoint maps, enabling relative transformation estimation.

formation rather than additional geometric descriptors.

C. Submap Correspondence Search

In multi-vehicle operations, establishing correspondences between maps is essential for global localization without an initial pose estimate. Vehicle i seeks to localize within vehicle j's map $\mathcal{M}_j$, leveraging its own map $\mathcal{M}_i$ to identify object correspondences.

We assume no temporal knowledge or trajectory information while constructing submaps. Instead, the environment map is partitioned into smaller submaps using a sliding window approach on $\mathcal{M}_i$, relying solely on the previously computed 3D object position estimates. We first compute the Mahalanobis distance of each landmark to the mean of the map and retain only the Ω^{th} percentile as inliers. Submaps are defined by a window size w, describing the length and width of each submap, and an overlap parameter α, controlling the step size between the submap centers. The previous work [3] used a naive approach to submap generation, defining subsets of objects by their identification numbers rather than leveraging their geometric locations. In this implementation, the submaps represent the true geometry of the overall environment map, where each submap contains up to N_{max} inlier objects whose Euclidean distances are closest to submap center. If fewer than N_{max} inliers exist within the window, the submap contains all available inliers.

We formulate the correspondence problem as a consistency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $v \in \mathcal{V}$ represents a candidate object correspondence, and edges $e \in \mathcal{E}$ connect geometrically consistent pairs. Geometric consistency is defined using pairwise distances; in the noise-free case, points $p_i, q_i \in \mathcal{M}_1$ are consistent with $p'_i, q'_i \in \mathcal{M}_2$ if $\|p_i - q_i\| = \|p'_i - q'_i\|$. In the presence of noise, edge weights are computed with

$$s(x) = \begin{cases} \exp\left(-\frac{x^2}{2\sigma^2}\right) & |x| \leq \epsilon \\ 0 & |x| > \epsilon \end{cases}, \tag{1}$$

where $x = \|p_i - q_i\| - \|p'_i - q'_i\|$, σ is the expected noise, and ϵ is tunable. These weights are stored in the affinity matrix A, with $A_{v,v} = 1$.

The densest geometrically consistent clique is found by solving

$$\max_{u \in \{0,1\}^{N_{\text{cand}}}} \frac{u^T A u}{u^T u}, \quad \text{s.t. } u_v u_{v'} = 0 \text{ if } A_{v,v'} = 0, \tag{2}$$

where N_{cand} is the total number of candidate correspondences. A threshold γ enforces a minimum distance between matched points to prevent duplicates caused by segmentation inconsistencies.

Each correspondence search returns the largest geometrically consistent set of object pairs, $\mathcal{S}$ (Fig. 3). Using Arun's method [39], we estimate the relative transformation $T_{ij} \in SE(3)$ that aligns vehicle i's reference frame to vehicle j's from these object correspondences, pruning transformations with pitch or roll exceeding θ_{RP}. We consider a correspondence successful if $|\mathcal{S}| > S_{\text{max}}$, balancing precision and recall.

IV. Experiments

We evaluate our performance on both the Båtvik seasonal dataset [3] and a custom Highbay dataset, which we collected for this work. The Highbay dataset is comprised of two challenging video sequences captured from both the nadir and oblique viewpoints collected in a large indoor flight testing facility. We compare the precision, recall, map size, and search time of VISTA to five baseline approaches.

A. Baselines

ORB [23] is a commonly used visual feature detection method, typically combined with RANSAC [40] for matching feature descriptors and rejecting outlier detections. We implement the ORB+RANSAC baseline implementation described in SOS-Match [3], as a place recognition module based on per-image descriptors. We detect a maximum of 500 ORB features in each keyframe, utilize a RANSAC backend to identify k-nearest neighbor correspondences ($k = 2$) between ORB features, and retain the feature matches that survive Lowe's ratio test [21].

We compare VISTA to two state-of-the-art learning-based feature matching approaches. We benchmark our results against LoFTR [26] as well as SuperPoint [24] feature detection front-end combined with SuperGlue [25] feature matching back-end. For both methods we use pretrained outdoor weights for the Båtvik seasonal dataset [3] and pretrained indoor weights for the Highbay datasets.

We also compare against AnyLoc [27], a VPR method that produces global image descriptors rather than local feature descriptors and is reported to perform robustly across challenging conditions such as varying environments, lighting, and viewpoints. Following the authors' best-performing configuration, we use DINOv2 (layer 31, facet *value*) with VLAD vocabulary (32 clusters) to generate global descriptors for each keyframe and compute cosine similarity between descriptor pairs for all images in the datasets.

The previously published work, SOS-Match [3] addressed localization under appearance changes induced by seasonal variation. SOS-Match outperformed state-of-the-art traditional and learning-based baseline approaches utilizing nadir

imagery, but has difficulties generalizing to imagery taken from the oblique camera angle.

B. Performance metrics

We evaluate VISTA using precision and recall. For each submap pair considered during correspondence search, we compute the IoU of their 3D point clouds. Submap pairs with sufficient overlap (IoU $> \theta_o$) are expected to yield a correct transformation. A transformation is considered correct if its roll and pitch are below θ_{RP} (ensuring dynamic feasibility), its yaw is below θ_Y, and its translation is below T_{max}, since all environments are aligned to a common reference frame. We define *hypothesized matches* as candidate transformations that appear correct to the algorithm based on having roll and pitch below θ_{RP}. Precision is the fraction of true correct transformations among all hypothesized matches. Recall is the fraction of true correct transformations for submap pairs with IoU $> \theta_o$ relative to the total number of overlapping submap pairs.

We vary the cardinality threshold, S_{max}, defining the minimum number of detected point correspondences between submaps, to produce precision and recall results for our approach and SOS-Match [3]. For the ORB+RANSAC baseline, we similarly vary the number of feature correspondences detected between keyframes. For SuperPoint+SuperGlue and LoFTR, we vary the minimum match confidence threshold for image pairs. For AnyLoc, we vary the minimum cosine similarity score.

We record the average runtime of one correspondence search step for VISTA compared to each baseline method. This correspondence search time is reflective of the time required to determine if a localization event has occurred in real-time settings. These results are computed using a AMD Ryzen Threadripper 3960X CPU with 128 GB RAM. The results for SuperPoint+SuperGlue, LoFTR and AnyLoc, which require a GPU for correspondence search, are computed using a NVIDIA RTX 3090 GPU.

Following prior work [3], we report only the correspondence search runtime, as our focus is on evaluating localization performance rather than segmentation efficiency. The segmentation and tracking components rely on pre-trained SAM [12] and SAM2 [13], whose reported inference time is approximately 1.4 seconds per image NVIDIA RTX 3090 GPU for the 32×32 grid of prompt points [41]. As in [3], we therefore omit detailed front-end runtime analysis and instead isolate correspondence search time as the metric most relevant to real-time localization performance.

C. Datasets

We leverage the Båtvik seasonal dataset [3], consisting of six 3.5 km UAV flights over a dense forest region recorded over the course of a year. This dataset presents a challenging scenario for long-term VPR due to the drastic appearance changes between drone flights induced by seasonal variation. Fig. 4 depicts sample nadir imagery from adjacent seasonal comparisons. These images highlight the significant appearance changes in the environment due sharpening of shadows, changes in lighting, foliage, and snow coverage between seasons. See [3] for further detail about this dataset.

We also test VISTA on a Highbay dataset to evaluate robustness to camera perspective variations. This dataset consists of two experiments, *Experiment 0* and *Experiment 1*, each containing imagery from both nadir and oblique viewpoints. Experiment 0 contains 18 flat pads with 2841/2684 tracked objects (nadir/oblique); Experiment 1 adds trees, poles, boxes, and backpacks for 36 total objects with 576/706 tracked objects. Notably, we follow different trajectories between the nadir and oblique viewpoint sequences, testing VISTA's robustness to objects observed from significantly different perspectives—a known failure mode for many VPR methods. Hyperparameters are summarized in Table I.

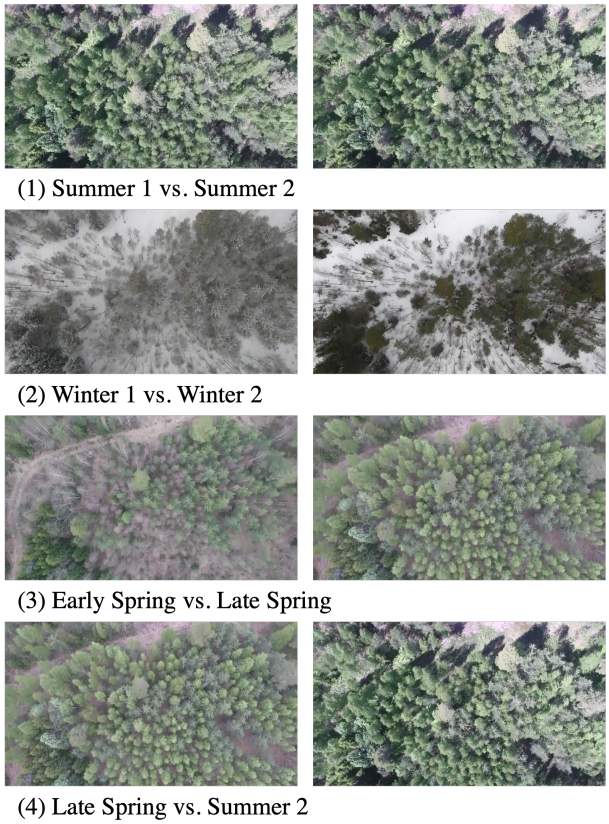

Fig. 4. Example imagery from the Båtvik seasonal dataset [3] demonstrating each of our four seasonal experiments. Adjacent seasonal imagery highlights the appearance changes induced by changes in foliage, sharpness of shadows, and snow coverage.

D. Seasonal Variation

We first evaluate VISTA in a scenario where two agents, both equipped with nadir-facing cameras, traverse the same trajectory under varying seasonal conditions. We assess performance across four experiments shown in Fig. 4: Summer

TABLE I
HYPERPARAMETERS

Exp.	Parameter	Value	Description
0 / 1	g_i	50 / 50	Image group batch size
0 / 1	Ω	[95/95] / [85/80]	Percentile inlier detections to keep [nadir/oblique]
0 / 1	w	2.0 / 2.0	Submap window size
0 / 1	α	1.0 / 1.0	Submap overlap parameter
0 / 1	n_{max}	50 / 50	Max. # of objects in submap
0 / 1	σ	0.05 m / 0.05 m	Pairwise consistency expected noise
0 / 1	ϵ	0.1 m / 0.1 m	Weighting function $s(x)$ cutoff
0 / 1	γ	0.1 m / 0.2 m	Min. distance between correspondences
0 / 1	S_{max}	4 / 4	Initial min. # correspondences
0 / 1	θ_o	0.667 / 0.667	Submap overlap threshold
0 / 1	θ_{RP}	10 deg / 6 deg	Roll/pitch threshold
0 / 1	θ_Y	30 deg / 30 deg	Yaw threshold
0 / 1	T_{max}	1.5 m / 1.5 m	Max. translation threshold

1 vs. Summer 2, Winter 1 vs. Winter 2, Early Spring vs. Late Spring, and Late Spring vs. Summer 2. Throughout these experiments, visual discrepancies progressively increase, posing more challenging scenarios for VPR methods.

We compare our results to the baseline methods using a camera projection to estimate the visible region at each keyframe, as described in [3]. We compute the IoU of the submaps being compared to determine if there is sufficient overlap for a valid correspondence. The precision and recall curves demonstrating the performance of VISTA compared to the baseline approaches is shown in Fig. 5.

VISTA outperforms all baseline methods in each of the four experiments. In the most challenging scenarios, with greater visual discrepancy between seasons, we can observe that the performance of all baseline methods significantly degrades. However, VISTA maintains consistent performance throughout the four scenarios and, as a result, shows a large performance improvement in these challenging cases. It is important to note that all of the baseline methods discard sections of the trajectory that fly over water due to the lack of visual features over these regions. The reported results for VISTA do not discard these sections of the trajectory, which further highlights the improved performance and robustness to appearance changes within the environment.

In Table II, we report the recall values at 100%, 90%, 80% precision for these seasonal variation scenarios. As precision decreases, recall will increase, however, it is impractical to utilize a system with very low precision, so we investigate the recall performance only when precision is high. As shown in Table II, VISTA achieves the highest recall values compared to the baseline approaches for all of the experiments and all precision thresholds with a maximum 42.8% improvement over the second best result as well as a maximum of 69% improvement over baseline methods at 100% recall.

E. Large Scale Simulation Environment

We first evaluate VISTA in a large-scale simulated environment [42] to verify consistent performance across environments of different scales. We perform this experiment due to the lack of publicly available, large-scale oblique-viewpoint aerial datasets and to assess scalability beyond the Highbay dataset environment. In Table III, we report recall performance at 100%, 90%, and 80% precision thresholds for both the nadir-only and nadir–oblique scenarios—the same evaluation protocol used for the Highbay datasets.

The results in Table III show that VISTA achieves near-perfect recall in the easy nadir-only configuration and slightly degraded recall in the more challenging nadir–oblique scenario. These outcomes closely match the Highbay dataset results (reported next), confirming that VISTA maintains consistent localization accuracy independent of environment scale. This consistency validates the use of the Highbay datasets for analyzing performance under oblique camera viewpoints.

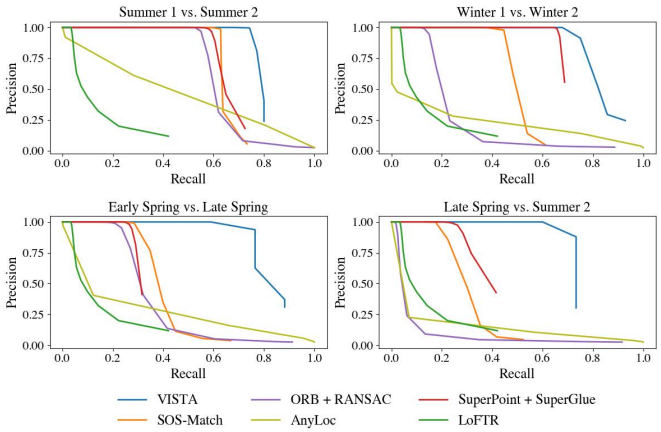

Fig. 5. Precision vs. recall curves with increasing visual discrepancy between flights induced by seasonal variation.

F. Oblique Camera Viewpoint

Building on the large-scale simulation results, we next evaluate VISTA on the Highbay datasets in a mixed-viewpoint scenario, where two agents—equipped with nadir and oblique camera configurations, respectively—operate within the same environment and aim to localize within each other's maps.

In Table IV, we present recall performance at 100%, 90%, and 80% precision thresholds. We first compare maps generated using only nadir-facing imagery, followed by those constructed using the nadir–oblique configuration. In the nadir-only scenario, VISTA achieves high recall (83.5%–90%) across all precision thresholds. In the more challenging nadir–oblique case, recall performance degrades relative to nadir-only, as expected.

We compare VISTA against five baselines under this oblique configuration (Table V). We report recall at 100%, 90%, and 80% precision, with map size and runtime comparisons in Table VI. VISTA outperforms all baselines at each precision threshold across both datasets, achieving 9–33% improvement over next-best results while maintaining maps only 0.03%–0.6% of baseline sizes. VISTA also outperforms all baselines except AnyLoc [27] in runtime; however, AnyLoc's speed comes at the cost of maps over 1200× larger, making inter-vehicle communication impractical.

Further, SOS-Match [3] struggles to maintain consistent feature tracks from the oblique viewpoint due to increased depth uncertainty and changes in object scale—and therefore

TABLE II

RECALL @ PRECISION, NADIR VS. NADIR, BÅTVIK SEASONAL DATASET, **BEST** RESULT FOR EACH PRECISION THRESHOLD HIGHLIGHTED, SECOND BEST RESULT UNDERLINED

Methods	Summer 1 vs. Summer 2			Winter 1 vs. Winter 2			Early Spring vs. Late Spring			Late Spring vs. Summer 2		
	R@100	R@90	R@80	R@100	R@90	R@80	R@100	R@90	R@80	R@100	R@90	R@80
SOS-Match [3]	25.6	62.9	62.9	12.4	45.4	46.5	17.5	31.0	33.9	12.7	20.9	23.5
ORB [23] + RANSAC [40]	46.9	55.9	57.3	9.9	15.5	16.5	11.1	24.5	26.6	1.2	2.4	2.8
AnyLoc [27]	0.0	2.9	11.7	0.0	0.03	0.06	0.0	1.7	3.8	0.0	0.9	1.8
SuperPoint [24] + SuperGlue [25]	52.8	60.3	61.3	62.6	66.9	67.5	22.4	28.0	29.1	17.8	28.5	30.5
LoFTR [26]	3.2	4.1	4.5	3.2	4.1	4.5	3.2	4.1	4.5	3.2	4.1	4.5
VISTA	**68.6**	**75.8**	**77.2**	**67.9**	**75.2**	**77.0**	**58.8**	**76.5**	**76.5**	**60.0**	**71.0**	**73.3**

TABLE III

RECALL @ PRECISION, LARGE-SCALE SIMULATED ENVIRONMENT

Method	AirSim [42] R@100	R@90	R@80
Nadir/Nadir	73.1	98.5	98.5
Nadir/Oblique	1.1	55.8	55.8

TABLE IV

RECALL @ PRECISION, HIGHBAY DATASET PERFORMANCE

	Experiment 1 R@100	R@90	R@80	Experiment 0 R@100	R@90	R@80
Nadir/Nadir	83.5	88.6	88.6	85.0	90.0	90.0
Nadir/Oblique	33.3	33.3	40.1	16.7	36.0	40.5

segmentation-mask size—used in its data association process. In this oblique-angle scenario, inconsistent object tracking leads to poor 3D object position estimates, degrading mapping performance. This in turn hinders accurate environment reconstruction and prevents the identification of submap correspondences necessary for robust global localization.

The robustness to occlusion and perceptual aliasing claimed in our abstract is substantiated by experimental evidence. SAM2's temporal tracking enables VISTA to maintain object correspondences through partial occlusions in the Highbay nadir–oblique experiments, where feature-based methods lose track and recall collapses (Table V). Similarly, VISTA's object-level geometric representation avoids the local-descriptor ambiguities that cause severe recall degradation for ORB, SuperGlue, LoFTR, and AnyLoc in the Båtvik seasonal dataset—where perceptual aliasing from homogeneous forest structure is severe (Table II). VISTA maintains 60–77% recall where ORB drops to 1.2%, demonstrating that geometric object configurations are less susceptible to aliasing than appearance-based features.

G. Ablation

Our auto-segmentation tracking pipeline leverages a segmentation model [12] to produce binary masks for all objects in each frame. We evaluate how alternative prompting strategies—points, bounding boxes, and masks—affect tracking performance by modifying our pipeline to specify the object of interest using each input type. Following [13], the video segmentation model accepts all three descriptors. Table VII reports recall at 100%, 90%, and 50% precision thresholds, with the lower thresholds included to accommodate degraded performance from the weaker prompting methods.

The results show that directly using segment masks provides the most reliable tracking signal. Masks preserve object geometry and contours, enabling stable pixel-level correspondences across scale changes, partial occlusions, and viewpoint shifts. Bounding boxes encode only coarse spatial extent and often include background or overlapping objects, producing ambiguous associations and drift. Point prompts remove shape information entirely, making them highly sensitive to localization noise or appearance changes.

As a result, bounding box prompting fails to achieve high recall or precision. Point prompting can reach high precision but with poor recall: on Experiment 0 it slightly exceeds our method at 100% precision but is surpassed at 90%. We attribute this to Experiment 0's simplicity (flat pads with limited object diversity), where geometric cues are less critical.

TABLE V

RECALL @ PRECISION, HIGHBAY NADIR VS. OBLIQUE, **BEST** AND SECOND BEST PERFORMANCE MARKED

Method	Experiment 1 R@100	R@90	R@80	Experiment 0 R@100	R@90	R@80
SOS-Match [3]	-	0.4	0.8	8.0	8.0	8.0
ORB [23] + RANSAC [40]	-	0.2	4.0	0.0	4.0	10.0
AnyLoc [27]	0.03	1.0	5.0	0.0	0.3	0.5
SuperPoint [24] + SuperGlue [25]	0.1	11.0	18.0	1.0	18.0	25.0
LoFTR [26]	0.3	3.0	4.0	0.03	6.0	7.0
VISTA	**33.3**	**33.3**	**40.1**	**17.0**	**36.0**	**41.0**

TABLE VI

MAP SIZE AND MATCH SEARCH TIME COMPARISON. **BEST** AND SECOND BEST PERFORMANCE ARE MARKED

Implementation	Map size (Mb)	Comparison runtime (s)	std
SOS-Match [3]	0.62	5.24	1.37
ORB [23] + RANSAC [40]	101.4	4663.3	15.3
AnyLoc [27]	708.6	**0.12**	**0.003**
SuperPoint [24] + SuperGlue [25]	1811.0	67.22	0.72
LoFTR [26]	1102.3	245.58	0.08
VISTA	**0.59**	1.08	0.67

In the more complex Experiment 1 environment, our mask-based method maintains superior localization performance.

TABLE VII

AUTO-SEGMENTATION TRACKER ABLATION

Prompting Method	Experiment 1 R@100	R@90	R@50	Experiment 0 R@100	R@90	R@50
Point	0.0	4.0	20.0	**33.3**	33.3	33.3
Bounding Box	-	-	5.3	0.0	0.0	39.6
Mask (ours)	**33.3**	**33.3**	**63.2**	16.7	**36.0**	**50.0**

V. DISCUSSION

This work highlights the power of combining deep learning models with sparse geometric front-end mapping methods to achieve our goal of localization across a team of vehicles operating in environments with significant appearance changes. Our auto-segmentation tracker identifies object segments using only image sequences as input to reconstruct 3D object-based maps. This offers two key advantages as opposed to previous work [3]. First, our approach eliminates the need for size descriptors, which we found to degrade quality, particularly in oblique view scenarios. Second, our approach removes the dependence on camera pose estimates in the tracking front-end, enhancing robustness by decoupling data association from potential pose estimation errors.

Our 3D object-based maps accurately preserve the environment geometry, enabling object correspondences to be identified between vehicle maps. We highlight that VISTA is not only capable of geometrically consistent mapping under extreme visual appearance changes induced by seasonal variation or different camera orientations, but also robust to changes in viewpoint, which we demonstrate in our Highbay experiments where vehicles reconstruct the environment following different trajectories. We ensure that our environment map is lightweight and runtime is low enough that VISTA can be used for localization or loop closure detection for teams of autonomous vehicles operating in real-time.

Our method does not attempt to identify when the vehicle has returned to a previously mapped location and does not explicitly handle duplicate objects within the map. Rather, we implicitly address these inconsistencies during the submap

correspondence search when computing the largest pairwise, geometrically consistent set of objects, we set a γ threshold, which prevents multiple points mapped within the threshold distance from being included in the maximum pairwise consistent set. In future work, we plan to utilize our method for loop closure detection within a SLAM framework. In addition, we plan to incorporate semantic information into our long-term localization framework to further disambiguate between submaps in these challenging monocular VPR scenarios under drastic environment appearance changes.

VI. Conclusion

This work presents VISTA, a monocular global localization framework designed for teams of autonomous vehicles operating within the same environment under extreme appearance changes. VISTA enables vehicles to localize within environment maps generated onboard other vehicles allowing for aligning reference frames to facilitate effective information sharing. This framework features lightweight object-based mapping and efficient correspondence search designed for open-set operation, making it well-suited for collaborative autonomous localization. Experiments demonstrate robust localization performance in the presence of seasonal variation, camera orientation differences, viewpoint variation, lighting changes, occlusions, and perceptual aliasing, highlighting the strengths of segmentation-based tracking and geometric search in unstructured environments.

References

[1] W. J. Wagner, I. Blankenau, M. DeLaTorre, A. Purushottam, and A. Soylemezoglu, "A robust localization solution for an uncrewed ground vehicle in unstructured outdoor gnss-denied environments," *ION GNSS+*, 2023.

[2] H. Yin, X. Xu, S. Lu, X. Chen, R. Xiong, S. Shen, C. Stachniss, and Y. Wang, "A survey on global lidar localization: Challenges, advances and open problems," *IJCV*, pp. 1–33, 2024.

[3] A. Thomas, J. Kinnari, P. C. Lusk, K. Kondo, and J. P. How, "Sos-match: segmentation for open-set robust correspondence search and robot localization in unstructured environments," in *IROS*. IEEE, 2024, pp. 5613–5620.

[4] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *T-RO*, vol. 28, no. 5, pp. 1188–1197, 2012.

[5] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," in *RA-L*, 2018.

[6] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *T-RO*, vol. 32, no. 1, pp. 1–19, 2015.

[7] 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," in *CVPR*, 2019, pp. 8092–8101.

[8] Y. Lin, J. Huang, and S. Lian, "Appearance-invariant 6-dof visual localization using generative adversarial networks," *arXiv preprint arXiv:2012.13191*, 2020.

[9] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *T-RO*, vol. 38, no. 4, 2022.

[10] S. Orhan, J. J. Guerrero, and Y. Baştanlar, "Semantic pose verification for outdoor visual localization with self-supervised contrastive learning," in *CVPR*, 2022, pp. 3989–3998.

[11] M. B. Peterson, Y. X. Jia, Y. Tian, A. Thomas, and J. P. How, "Roman: Open-set object map alignment for robust view-invariant global localization," *arXiv preprint arXiv:2410.08262*, 2024.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[13] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[14] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison, "Elasticfusion: Dense slam without a pose graph." in *RSS*, vol. 11. Rome, Italy, 2015, p. 3.

[15] V. Yugay, Y. Li, T. Gevers, and M. R. Oswald, "Gaussian-slam: Photo-realistic dense slam with gaussian splatting," 2023.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *CACM*, vol. 65, no. 1, pp. 99–106, 2021.

[17] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ToG*, vol. 32, no. 6, pp. 1–11, 2013.

[18] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *T-RO*, vol. 32, no. 6, p. 1309–1332, Dec. 2016.

[19] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *T-RO*, vol. 35, no. 4, pp. 925–938, 2019.

[20] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *RA-L*, vol. 4, no. 1, pp. 1–8, 2018.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[22] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.

[23] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*. Ieee, 2011, pp. 2564–2571.

[24] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *CVPR workshops*, 2018, pp. 224–236.

[25] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *CVPR*, 2020, pp. 4938–4947.

[26] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021, pp. 8922–8931.

[27] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *RA-L*, 2023.

[28] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *TPAMI*, vol. 44, no. 7, pp. 3523–3542, 2021.

[29] M. Javanmardi, E. Javanmardi, Y. Gu, and S. Kamijo, "Towards high-definition 3d urban mapping: Road feature-based registration of mobile mapping systems and aerial imagery," *Remote Sensing*, vol. 9, no. 10, p. 975, 2017.

[30] G. Kim, S. Choi, and A. Kim, "Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments," *T-RO*, vol. 38, no. 3, pp. 1856–1874, 2021.

[31] O. Pink, "Visual map matching and localization using a global feature map," in *CVPR workshops*. IEEE, 2008, pp. 1–7.

[32] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak *et al.*, "Present and future of slam in extreme environments: The darpa subt challenge," *T-RO*, vol. 40, pp. 936–959, 2023.

[33] Y. Tian, K. Liu, K. Ok, L. Tran, D. Allen, N. Roy, and J. P. How, "Search and rescue under the forest canopy using multiple uavs," *IJRR*, vol. 39, no. 10-11, pp. 1201–1221, 2020.

[34] T. Pritchard, S. Ijaz, R. Clark, and B. B. Kocer, "Forestvo: Enhancing visual odometry in forest environments through forestglue," *RA-L*, 2025.

[35] P. S. Szczepaniak, "Interpretation of image segmentation in terms of justifiable granularity," in *Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part I 14*. Springer, 2015, pp. 638–648.

[36] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao, "Semantic-sam: Segment and recognize anything at any granularity," 2023.

[37] F. Dellaert, M. Kaess *et al.*, "Factor graphs for robot perception," *Found. and Trends® in Robot.*, vol. 6, no. 1-2, pp. 1–139, 2017.

[38] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," *Georgia Institute of Technology, Tech. Rep*, vol. 2, p. 4, 2012.

[39] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *TPAMI*, no. 5, pp. 698–700, 1987.

[40] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *CACM*, vol. 24, no. 6, pp. 381–395, 1981.

[41] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," 2023.

[42] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*. Springer, 2018, pp. 621–635.