

# ShelfAware: Real-Time Visual-Inertial Semantic Localization in Quasi-Static Environments with Low-Cost Sensors

Shivendra Agrawal, Jake Brawer, Ashutosh Naik, Alessandro Roncone, and Bradley Hayes

**Abstract**—Many indoor workspaces are *quasi-static*: their global geometric layout is stable, but local semantics change continually, producing repetitive geometry, dynamic clutter, and perceptual noise that defeat standard vision-based localization. We present *ShelfAware*, a semantic particle filter for robust global localization that treats scene semantics as *statistical evidence over object categories* rather than fixed quantity landmarks. ShelfAware fuses a depth likelihood with a category-centric semantic similarity and uses a precomputed bank of semantic viewpoints to perform *inverse semantic proposals* inside Monte Carlo Localization (MCL), yielding fast, targeted hypothesis generation on low-cost, vision-only hardware. To demonstrate perception-agnostic scalability, we evaluate ShelfAware across two domains. In a rigorously controlled mock retail environment, ShelfAware achieves a 97% global localization success rate, maintaining the highest tracking success (66%) across cart, wearable, and dynamic occlusion conditions. Furthermore, in a 3,500 sq. ft. operational grocery store leveraging an open-vocabulary vision pipeline, ShelfAware significantly outperforms both geometric and fixed-quantity semantic baselines. By modeling semantics distributionally and leveraging inverse proposals, ShelfAware resolves geometric aliasing, providing an infrastructure-free building block for mobile and assistive robots in dynamic real-world environments.

**Index Terms**—Localization, Semantic Scene Understanding, SLAM

## I. INTRODUCTION

MANY real-world indoor environments, such as retail stores and warehouses, are *quasi-static*: their global geometric layout (e.g., walls, shelving units) remains permanent over long horizons, but their local semantic contents change continually. In these settings, traditional geometry-based particle filters like Adaptive Monte Carlo Localization (AMCL) [1] suffer from severe perceptual aliasing. The combination of visually repetitive aisle geometry, dynamic clutter, and the noisy observations typical of low-cost wearable or mobile sensors routinely violates static map assumptions and causes rapid particle impoverishment [2]. Consequently, robust, start-anytime global localization in these GPS-denied environments remains a critical open problem for autonomous and assistive systems.

To address these challenges, we present ShelfAware (Fig. 1), a semantic particle filter tailored for robust global localization in quasi-static environments. Unlike prior approaches

Manuscript received: December 9, 2025; Revised: March 4, 2026; Accepted: March 30, 2026.

This paper was recommended for publication by Editor Ayoung Kim upon evaluation of the Associate Editor and Reviewers' comments.

The authors are with the Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309 USA (e-mail: shivendra.agrawal@colorado.edu).

Digital Object Identifier (DOI): see top of this page.

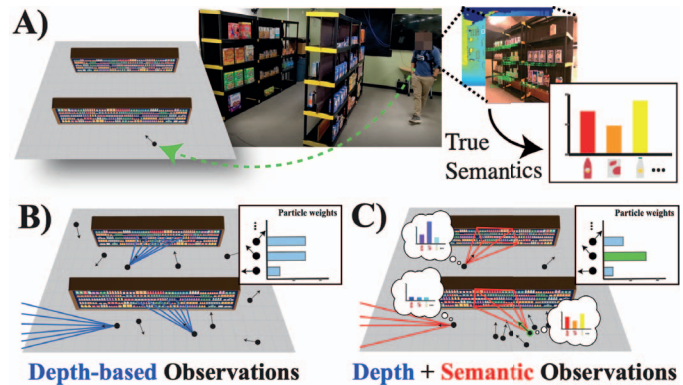


Fig. 1: An overview of ShelfAware. A) A retail-like environment. B) Depth-based observation models in particle filtering rely solely on geometric features, which are ambiguous in long, repetitive aisles and lead to weak particle discrimination. C) ShelfAware injects particles based on semantic cues, enabling more distinctive and robust particle weighting combined with the depth observation model and improved global localization in retail-like environments.

that treat semantics as fixed, discrete landmarks [3]–[6], ShelfAware models semantic features as probabilistic distributions over object counts and spatial arrangements. This representation captures the intrinsic variability of real-world environments, preserving the statistical structure necessary for stable observation-driven localization even as local object configurations continuously fluctuate.

We evaluate ShelfAware across two domains: a semantically dense mock retail environment enabling rigorous motion-capture ground truth, and a 3,500 sq. ft. operational grocery store demonstrating open-world scalability. Experiments on consumer-grade wearable and cart-mounted sensors highlight the system's robustness for practical deployment across both autonomous service robots and assistive devices.

Our core technical contributions are twofold:

- A semantic mapping and observation design that encodes object collections as statistical distributions over counts and arrangement, providing inherent robustness to semantic perturbations and flux.
- A real-time, inverse observation model-based particle filter that leverages this representation on low-cost, vision-based hardware to efficiently propose and evaluate high-quality global pose hypotheses.

## II. RELATED WORK

### A. Vision and Semantics-Based Localization in Quasi-Static Environments

Particle filter-based localization methods such as Monte Carlo Localization (MCL) and its adaptive variant AMCL [1] remain standard for onboard robot localization due to their scalability and integration with modern navigation frameworks [7]. However, their reliance on static geometric maps and feature-rich depth data makes them brittle in quasi-static or dynamic environments where geometry is repetitive or transient. This limitation has driven research toward semantic localization, where environmental understanding extends beyond geometry to include object-level cues and contextual information [8].

Earlier work in semantic localization augments geometric maps with explicit, object- or region-level labels that serve as long-term landmarks for place recognition and drift reduction [4]–[6]. Text-based cues, such as signage or packaging, have also proven effective as distinctive features in structured indoor spaces [3]. More recent efforts adopt implicit semantic representations, encoding spatial and semantic structure jointly in learned neural fields [9], enabling continuous observation models compatible with particle filtering. Yet, both explicit and implicit methods often assume stable semantic identities, assumptions that rarely hold in quasi-static domains like warehouses or retail stores, where local semantics fluctuate even when global geometry remains constant.

Despite these advances, few approaches explicitly address semantic volatility: the gradual yet continual change in object distributions and arrangements that characterizes quasi-static environments. Traditional semantic-SLAM systems [5], [10] and semantic visual positioning frameworks [11], [12] demonstrate the promise of integrating semantics into localization, but their models typically treat detected landmarks as fixed or sparsely varying entities. This mismatch between model assumptions and real-world semantic drift leads to degraded performance in settings with restocking, occlusions, or partial observability. Recent deep visual localization methods [13], [14] demonstrate impressive accuracy but require server-grade compute, precluding real-time use on wearable edge devices. Furthermore, while modern visual SLAM [15] excels at local tracking, it typically requires extensive exploration to achieve global localization against a prior map, limiting the instantaneous recovery needed for on-demand assistance.

### B. Applications in Assistive and Human-Centered Robotics

Quasi-static indoor environments such as retail stores also appear prominently in assistive and human-centered robotics. Many assistive navigation systems for people with visual impairments depend on reliable global localization to support guidance, object retrieval, or wayfinding [16]–[18]. However, prior systems often sidestep this challenge, relying instead on environmental instrumentation (e.g., RFID tags [17], Bluetooth beacons) or assuming that the user is already localized within a specific aisle or region [19]. Recent conversational and multimodal assistance frameworks [20], [21] have improved interaction but still rely on external positioning aids.

ShelfAware complements this body of work by targeting the underlying localization problem directly, enabling global localization without external infrastructure, and doing so in a semantically dynamic environment representative of those faced by both assistive and service robots.

In contrast to prior methods that treat semantics as static landmarks, ShelfAware models them as statistical distributions over object counts and arrangements, enabling localization that is both robust to semantic flux and compatible with low-cost visual sensing. This probabilistic treatment of semantics situates ShelfAware within the broader context of semantic particle filtering, extending its applicability beyond assistive scenarios to general quasi-static domains.

## III. OUR APPROACH

The objective of our proposed method is to achieve robust global localization in quasi-static indoor environments. These environments, which include warehouses, retail spaces, and laboratories, challenge conventional geometric localization methods due to visually repetitive structures, dynamic occlusions, and semantic drift [2]. ShelfAware addresses these challenges by fusing geometric and semantic cues within a probabilistic particle filtering framework tailored for vision-based sensing, and deployable on wearable or compact platforms that preclude LiDAR or wheel odometry.

### A. Semantic Particle Filter Overview

ShelfAware builds on the MCL framework [1], augmenting standard depth likelihoods with a probabilistic semantic observation model that remains informative under semantic variability. Each particle represents a pose hypothesis  $\mathbf{x}_t^{(i)}$  with weight  $w_t^{(i)}$ , updated via motion, geometric (depth), and semantic observations (Fig. 4).

Unlike approaches that treat detected objects as fixed landmarks, ShelfAware models the semantic state of the environment as distributions over class counts and coarse spatial arrangement. At runtime, the system forms a compact semantic observation vector from the live RGB-D frame and compares it to expected semantic signatures derived from a hybrid map (Sec. III-B). The resulting semantic similarity acts both as (i) a *forward* observation likelihood for weighting particles and (ii) a query metric for an *inverse* semantic model that proposes high-quality pose hypotheses when global localization or recovery is needed (Sec. III-E).

The joint observation model factors geometric and semantic likelihoods as such:

$$p(\mathbf{z}_t | \mathbf{x}_t, m) = \eta p_d(\mathbf{z}_t^d | \mathbf{x}_t, m_d) p_s(\mathbf{z}_t^s | \mathbf{x}_t, m_s), \quad (1)$$

where  $\eta$  is a normalizer,  $m_d$  and  $m_s$  are the geometric and semantic maps respectively,  $\mathbf{z}_t = (\mathbf{z}_t^d, \mathbf{z}_t^s)$  are the depth and semantic observations (Secs. III-C, III-D). This factorization assumes conditional independence: given pose  $\mathbf{x}_t$ , depth  $\mathbf{z}_t^d$  and semantic  $\mathbf{z}_t^s$  measurements have independent noise.

## B. Semantic Mapping

ShelfAware maintains a hybrid map combining geometric structure with semantic information. First, we construct a standard 2D occupancy grid map of traversable space using GMapping [22], with  $10 \times 10$  cm resolution. This resolution supports precise ray casting for the depth observation model within the MCL framework and aligns with the accuracy needed for potential downstream manipulation tasks [19].

Second, we build a semantic map overlaid on the occupancy grid. The semantic map discretizes the volume into  $20 \times 20$  cm cells in  $(x, y)$  and 30 cm in  $z$  to balance computational efficiency with the physical scale of objects and the environment. Each voxel stores a running distribution over observed object *classes* and their *counts*, yielding a coarse, distributional representation that captures typical arrangements without assuming fixed landmark identities.

To demonstrate that our system can work with different vision pipelines, we populate this semantic map using two distinct pipelines, depending on the evaluation environment:

a) *Pipeline A: Supervised Closed-Set Semantics:* For rigorous quantitative benchmarking in a controlled mock environment (Sec. IV-A) with fixed and known number of classes, we utilize a YOLOv9 detector fine-tuned on the SKU-110K dataset [23] to extract generic “shelf item” bounding boxes. A ResNet50-based classifier then assigns these detections to one of  $C = 14$  manually defined application-level classes.

b) *Pipeline B: Self-Supervised Open-Set Semantics:* To demonstrate zero-shot scalability for real-world deployment in a 3,500 sq. ft. operational grocery store, we eliminate manual annotation entirely. Using the same class-agnostic detector, we extract a 768-dimensional vector for each bounding box crop via a pre-trained DINOv3 Vision Transformer. We then cluster these features using K-Means, automatically defining  $C = 60$  (obtained via grid-search) semantic categories (Fig. 2:right).

Regardless of the chosen perception pipeline, detected objects are projected into the map frame using the RGB-D depth channel and camera calibration.

Given pixel coordinates  $\tilde{\mathbf{u}} = [u, v, 1]^\top$  at the bounding box’s median depth  $Z_c$ , the 3D world position is  $\mathbf{X}_w = \mathbf{t}_{wc} + Z_c \mathbf{R}_{wc} \mathbf{K}^{-1} \tilde{\mathbf{u}}$ , where  $\mathbf{K}$  is the camera intrinsic matrix, and  $\mathbf{R}_{wc}, \mathbf{t}_{wc}$  denote the camera-to-world transform.

To mitigate depth noise and vision errors, we refine each estimated product position along the camera ray using Bresenham’s line algorithm [24] until it aligns with the nearest occupied occupancy map cell, preventing minor misalignments. Figure 2 visualizes the resulting semantic layer overlaid on the occupancy grid; for clarity, the plot shows only the dominant class per cell, while the map stores full per-class count distributions.

## C. Semantic Observation Vector and Semantic Similarity

ShelfAware generates a semantic observation  $\mathbf{z}_t^s$  from the live camera view (Fig. 3). This observation concatenates three sub-vectors: (i) a class-count vector  $\mathbf{v}_c$  (what is present), (ii) a mean range vector  $\mathbf{v}_d$  (how far), and (iii) a mean bearing

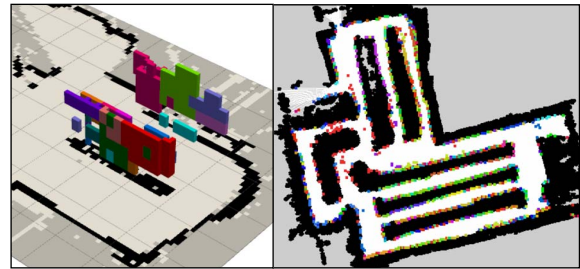


Fig. 2: The semantic map overlaid on the occupancy grid. Each voxel stores a distribution over object class counts. Ray casting on this semantic layer yields the expected semantic vector  $\mathbf{v}_{\text{sem}}$  comprising class counts, distances, and angles. *Left:* Mock store, *Right:* Real Store.

vector  $\mathbf{v}_\theta$  (in which direction), each aggregated over detections for the classes visible at time  $t$ .

To compare live and expected semantic observations, we define a composite similarity

$$S(\mathbf{z}^s, \hat{\mathbf{z}}^s) = \alpha S_{\text{counts}} + \beta S_{\text{distance}} + \gamma S_{\text{angle}}, \quad (2)$$

with  $\alpha, \beta, \gamma \geq 0$  and  $\alpha + \beta + \gamma = 1$ . This score is used both as (i) a forward semantic likelihood  $p_s(\mathbf{z}_t^s | \mathbf{x}_t, m_s)$  in (1) and (ii) a query metric for the inverse model (Sec. III-E).

a) *Counts.:* We normalize  $\mathbf{v}_c$  to form a categorical distribution over observed classes and compare it with the expected distribution via Jensen–Shannon divergence (JSD). Defining  $P$  and  $Q$  as the normalized count distributions and  $M = \frac{1}{2}(P + Q)$ , we set  $S_{\text{counts}} = 1 - \text{JSD}$  with  $\text{JSD}(P \| Q) = \sqrt{\frac{1}{2} \sum_{i=1}^C \left( P(i) \log \frac{P(i)}{M(i)} + Q(i) \log \frac{Q(i)}{M(i)} \right)}$ . This choice is symmetric, bounded, and robust to sparsity in per-frame detections. By comparing normalized distributions rather than absolute counts, this JSD formulation provides inherent robustness to perception failures. Transient false negatives (e.g., missed bounding boxes due to motion blur) or false positives do not catastrophically invalidate the observation, but rather smoothly degrade the similarity score, preventing abrupt particle depletion.

b) *Distances and Angles.:* The spatial components  $\mathbf{v}_d$  and  $\mathbf{v}_\theta$  are compared using L2 distances. For unbounded ranges, we map the L2 error  $d_{\text{distance}}$  to  $S_{\text{distance}} = 1/(1 + d_{\text{distance}})$ . For angles, which lie within the camera field-of-view (FOV), we use  $S_{\text{angle}} = 1 - (d_{\text{angle}}/\text{FOV})$ .

## D. Depth Observation Model

To incorporate geometric information, we synthesize a 2D laser scan from a central horizontal band of the depth image. The depth observation at time  $t$  is  $\mathbf{z}_t^d = \{z_t^{d,(k)}\}_{k=1}^K$ . Given the expected range  $\hat{z}^{(k)}(\mathbf{x}_t, m_d)$  for the  $k$ -th beam (via occupancy-map ray casting), we use the beam end-point mixture [1]:

$$p_{\text{hit}}(z_t^{d,(k)} | \mathbf{x}_t, m_d) = \mathcal{N}(z_t^{d,(k)}; \hat{z}^{(k)}(\mathbf{x}_t, m_d), \sigma_{\text{hit}}^2), \quad (3)$$

$$p_{\text{short}}(z_t^{d,(k)} | \mathbf{x}_t, m_d) = \lambda e^{-\lambda z_t^{d,(k)}} \mathbf{1}_{[0, \hat{z}^{(k)}(\mathbf{x}_t, m_d)]}(z_t^{d,(k)}), \quad (4)$$

$$p_{\text{max}}(z_t^{d,(k)}) = \delta(z_t^{d,(k)} - z_{\text{max}}), \quad p_{\text{rand}}(z_t^{d,(k)}) = 1/z_{\text{max}}, \quad (5)$$

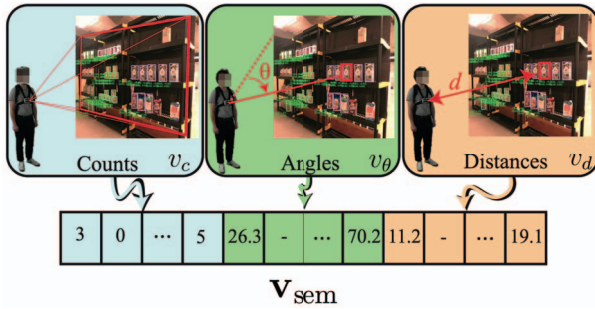


Fig. 3: Semantic vector  $\mathbf{v}_{\text{sem}} = [\mathbf{v}_c, \mathbf{v}_\theta, \mathbf{v}_d]$ . The count vector  $\mathbf{v}_c$  captures the number of items detected in each class;  $\mathbf{v}_\theta$  and  $\mathbf{v}_d$  capture mean relative bearings and ranges for those.

$$p_d(z_t^{d,(k)} | \mathbf{x}_t, m_d) = w_h p_{\text{hit}} + w_s p_{\text{short}} + w_m p_{\text{max}} + w_r p_{\text{rand}}, \quad (6)$$

where  $m_d$  represents the 2D occupancy map,  $z_{\text{max}} = 6.0$  m is the maximum reliable sensor range,  $\sigma_{\text{hit}} = 2.0$  is the standard deviation of the expected depth, and  $\lambda = 0.1$  is the intrinsic decay parameter for short readings. The empirically tuned mixture weights ( $w_h = 0.85$ ,  $w_s = 0.05$ ,  $w_m = 0.05$ ,  $w_r = 0.05$ ) define the relative probability of a valid hit, a short reading, a max-range reading, and a random noise reading, respectively, and are constrained to sum to one. Assuming conditional independence across beams, the joint likelihood for  $\mathbf{z}_t^d$  is

$$p_d(\mathbf{z}_t^d | \mathbf{x}_t, m_d) = \prod_{k=1}^K p_d(z_t^{d,(k)} | \mathbf{x}_t, m_d). \quad (7)$$

We compute  $\hat{z}^{(k)}(\mathbf{x}_t, m_d)$  efficiently using CDDT-accelerated ray casting [25], which runs on the CPU and leaves GPU resources available for the semantic perception pipeline.

### E. Localization with an Inverse Semantic Model

ShelfAware integrates the depth and semantic models within an MCL filter, maintaining weighted particles  $\{(\mathbf{x}_t^{(i)}, \mathbf{w}_t^{(i)})\}_{i=1}^N$  over the pose belief. The method's key innovation is an *inverse semantic model* that proposes high-quality pose hypotheses directly from live semantic observations, enabling fast *global localization* and recovery from tracking failures without special-case handling.

a) *Offline Pre-computation*: We precompute expected semantic observations  $\hat{\mathbf{z}}^s(\mathbf{x}, m_s)$  for discretized poses  $\mathbf{x} = (x, y, \theta)$  over the free space (10cm cells; 36 orientation bins). For each pose, we ray cast the 3D semantic map (Sec. III-B) to obtain the expected semantic vector. All vectors are cached in a hashmap (76 MB for our store environment). We also build a reverse index `class_to_poses` (2.1 MB) mapping each product class to the set of poses from which that class is visible, enabling efficient candidate pruning at runtime.

b) *Online Localization Loop*: Algorithm 1 summarizes the online filter. Particles are propagated using VIO-based motion. Each iteration forms a live semantic vector  $\mathbf{z}_t^s$  and estimates an expected semantic view  $\hat{\mathbf{z}}_t^s$  at the current pose estimate  $\hat{\mathbf{x}}$ . A semantic consistency check compares  $S(\mathbf{z}_t^s, \hat{\mathbf{z}}_t^s)$  against a threshold and verifies that the count subvector has sufficient mass ( $\|\mathbf{z}_t^{s,\text{count}}\|_1 > \tau_\kappa$ ). If the check fails, we query

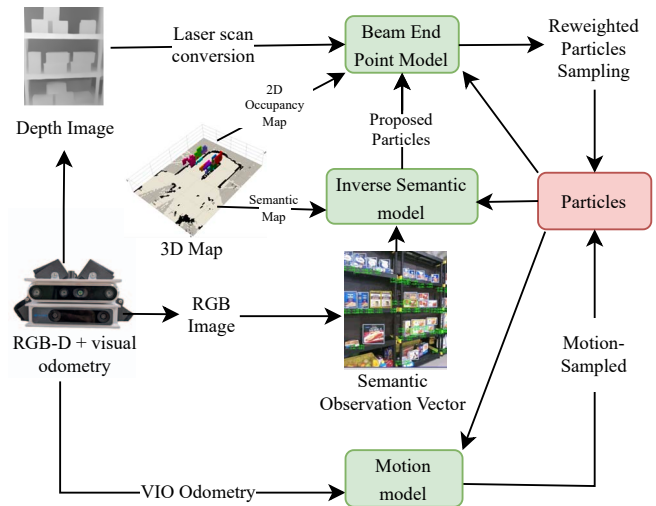


Fig. 4: Data-flow diagram for ShelfAware. The semantic particle filter fuses depth likelihood with semantic likelihood and uses an inverse semantic model to propose high-quality particles for global localization and recovery.



Fig. 5: ShelfAware hardware. A lightweight two-camera system with a 3D-printed mount was used throughout our experiments (top). This design allowed evaluation across a wearable chest mount (left) and a cart-mounted setup (right).

the inverse model: (i) use `class_to_poses` to form a candidate set (union over observed classes), (ii) score candidates by  $S$  (Eq. 2), (iii) inject particles at the top- $k$  poses, and (iv) reweight the entire set with both depth (Sec. III-D) and forward semantic likelihoods (Eq. 1). Otherwise, we update using only the depth likelihood for efficiency. This procedure enables rapid convergence from unknown initial conditions and robust recovery from drift.

### F. Hardware and Form Factors

To meet the constraints of low-profile, wearable, or cart-mounted platforms, ShelfAware uses compact vision sensors: (i) an Intel RealSense D455 RGB-D camera ( $\sim 103$  g) for color and depth, and (ii) an Intel RealSense T265 VIO camera ( $\sim 60$  g) for odometry. Both are housed in a single 3D-printed mount and connected via USB to a Dell G15 laptop. This configuration supports two form factors: a chest-mounted wearable (laptop in a backpack) and a cart-mounted

---

**Algorithm 1** Online Localization with ShelfAware

---

**Input:** Sensor data: image  $i_t$ , depth  $d_t$ , odometry  $o_t$   
Pre-computed map data: semantic vectors  $\hat{\mathbf{z}}^s(\mathbf{x})$ ,  
**Output:** Updated particle set  $P_{t+1}$   
**Initialize:** Particle set  $P_t$  sampled uniformly over free space

- 1: **while** True **do**
- 2:    $P_t \leftarrow \text{MotionModel}(P_t, o_t)$
- 3:    $z_t^s \leftarrow \text{GenerateSemanticVector}(i_t, d_t)$
- 4:    $\hat{x} \leftarrow \text{ExpectedPose}(P_t, W_t)$
- 5:    $\hat{z}_t^s \leftarrow \text{CalculateExpectedSemanticObs}(\hat{x})$
- 6:   // *Semantic consistency and information sufficiency*
- 7:    $\text{sim} \leftarrow \text{CalculateSemanticSimilarity}(z_t^s, \hat{z}_t^s)$
- 8:   **if**  $\text{sim} < \tau_{\text{sim}}$  **and**  $\|z_t^s\|_1 > \tau_{\kappa}$  **then**
- 9:      $C \leftarrow \text{GetObservedClasses}(z_t^s)$
- 10:      $\text{Candidates} \leftarrow \bigcup_{c \in C} \text{class\_to\_poses}[c]$
- 11:      $\text{Scores} \leftarrow [S(z_t^s, \hat{\mathbf{z}}^s(\mathbf{x})) \text{ for } \mathbf{x} \in \text{Candidates}]$
- 12:      $\text{TopPoses} \leftarrow \text{GetTopK}(\text{Candidates}, \text{Scores}, k)$
- 13:      $P_t \leftarrow \text{InjectParticles}(P_t, \text{TopPoses})$
- 14:      $W_{\text{depth}} \leftarrow \text{DepthObservationModel}(P_t, d_t)$
- 15:      $W_{\text{sem}} \leftarrow \text{SemanticObservationModel}(P_t, i_t, d_t)$
- 16:      $W_t \leftarrow W_{\text{depth}} \odot W_{\text{sem}} \quad \triangleright \text{element-wise product}$
- 17:   **else**
- 18:      $W_t \leftarrow \text{DepthObservationModel}(P_t, d_t)$
- 19:   **end if**
- 20:    $W_t \leftarrow \text{NormalizeWeights}(W_t)$
- 21:    $P_t \leftarrow \text{Resample}(P_t)$
- 22:    $P_{t+1} \leftarrow P_t$
- 23: **end while**

---

setup (Fig. 5). As discussed in Secs. III-B–III-E, the system exploits CPU-based CDDT ray casting [25] to reserve GPU resources for real-time semantic perception, enabling practical deployment without LiDAR or wheel encoders.

#### IV. EXPERIMENTAL EVALUATIONS

We evaluate ShelfAware’s ability to perform reliable, real-time global localization in quasi-static, semantically dense, and GPS-denied environments using compact vision sensors. Specifically, we assess its robustness to dynamic occlusions and sparse semantics, and its form factor sensitivity across wearable and cart-mounted configurations. We adopt these form factors to reflect practical compute constraints and to stress-test the method under depth-camera noise (lower scan frequency/point density than LiDAR) [26], intentionally avoiding the usability barriers of handheld devices [27] and the social stigma of head-mounted systems [28]. Ultimately, these evaluation goals reflect the start-anytime and recover-anytime operation demanded by practical deployments and shared-control use cases [29]–[31].

##### A. Experimental Setup

We conduct our controlled evaluation in a mock grocery store stocked with 150 products grouped into 14 object categories across nine shelves and three aisles. All experiments ran on a Dell G15 laptop (Intel Core i7-11800H (2.3 GHz); NVIDIA RTX 3060, 6GB VRAM; 32GB RAM) using the

same vision-only sensor suite described in Sec. III-F: an Intel RealSense D455 RGB-D camera and a RealSense T265 VIO camera (Fig. 5). Ground-truth 2D poses were obtained using an OptiTrack motion-capture system. We aligned the OptiTrack frame to the map frame offline by time-synchronizing mapping trajectories and solving for the rigid transform via RANSAC [32], enabling direct comparison of estimated and true poses. RGB-D streams were recorded at 30Hz and VIO at 200Hz (from the T265 IMU). This evaluation setting stresses depth-camera uncertainties that degrade purely geometric localization methods [26].

We considered four conditions to measure and characterize the robustness of the proposed method:

- 1) **Cart:** Cameras mounted on a cart (Fig.5-right) [33].
- 2) **Wearable:** Cameras chest-mounted; laptop carried in a backpack (Fig.5-left), stressing odometry noise introduced by gait.
- 3) **Dynamic Obstacles:** A person intermittently walked in front of the cameras to occlude vision.
- 4) **Sparse Semantics:** To mimic depleted stock, we randomly removed 25% and 50% of products from shelves, reducing category signal.

We collected 5 trajectory sequences per condition (20 total, S1–S20), each  $\sim 40$  s in duration. Between mapping and localization runs, we perturbed 20% of shelf contents to induce semantic drift. Unless stated otherwise, particles were initialized *uniformly over free space* to require true global localization [16], [33].

##### B. Perception Pipeline Configurations

We utilize the two perception pipelines introduced in Sec. III-B. For the mock store (Pipeline A), a YOLOv9 detector (fine-tuned on the SKU-110K dataset [23]) proposes generic bounding boxes, which a ResNet50 classifier maps into 14 object categories (trained on 5,699 environment-specific frames). Grouping thousands of underlying SKU types into a small, fixed category set provides robustness against local inventory churn. Conversely, for the real-world operational store (Pipeline B), we combine YOLOv9’s class-agnostic bounding box proposals with a zero-shot DINOv3 Vision Transformer and K-Means clustering. This open-vocabulary approach automatically extracts 60 semantic categories from the generic detections in a purely self-supervised manner, requiring zero environment-specific training or manual annotation to adapt to new inventory.

##### C. Baselines and Metrics

**Baselines.** We compare against three methods: (i) the standard ROS implementation of AMCL [7], (ii) an ablated ShelfAware variant that removes the semantic likelihood, yielding a pure geometric MCL baseline [1], and (iii) a Fixed Quantity Landmark (FQL) baseline. The FQL baseline represents standard explicit semantic localization approaches [3], [4] by treating detected object classes as having fixed, deterministic counts. For fairness, all methods used **1,500 particles**, identical VIO odometry and RGB-D stream.

**Metrics.** Following established conventions in the literature, we evaluate ShelfAware via:

- **Global localization success.** A trial is a success if localization convergence occurs within the first 95% of the trajectory and remains converged until the end.
- **Convergence time (s).** Time from start to the beginning of the final convergence for successful trials.
- **Tracking ATE.** Absolute Trajectory Error (translation/rotation RMSE) computed after final convergence until end-of-sequence.

**Convergence criterion.** Anthropometric studies indicate a standing reach of  $\sim 0.7\text{--}0.9$  m [34]; thus, localization errors below 0.7 m ensure targets remain within the reachable workspace. We declare convergence when the estimated pose is within 0.7 m translation and  $\pi/4$  rad rotation of ground truth and stays within that threshold. We tuned semantic similarity weights by grid search and report results for  $\alpha = 0.4, \beta = 0.4, \gamma = 0.2$  (Sec.III-C).

#### D. Results: Global Localization

Table I (four conditions, 25 trials each) summarizes success, convergence time, and RMSE. ShelfAware achieved a **97%** overall success rate across 100 trials, versus **15%** for MCL and **10%** for AMCL. The mean time-to-convergence across all setups was **1.07s**. ShelfAware obtained the lowest translational RMSE in three out of four conditions.

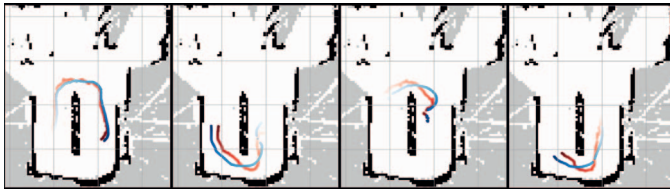


Fig. 6: Examples of ground truth pose in blue and our estimated pose in red. Lighter to darker denotes the temporal progression of the trajectories. Each grid cell is  $2\text{m} \times 2\text{m}$ .

#### E. Results: Tracking and Robustness

To evaluate continuous tracking performance, we measure the Tracking Success Rate (T%), defined as the percentage of trials that converge and remain strictly within the convergence thresholds until the end. Table II reports the T% and the average tracking RMSE (T-RMSE) across all conditions.

ShelfAware achieved an overall tracking success rate of **66%**, outperforming the deterministic FQL baseline (59%) and vastly exceeding the purely geometric MCL and AMCL methods (10%). Importantly, the system demonstrated particular resilience against the two primary environmental stressors: dynamic occlusions and sparse semantics. Qualitative trajectories (Fig.6) show rapid correction in repetitive aisles and robustness to occlusion bursts. These results mirror the global-localization advantage, demonstrating that probabilistic semantic tie-breaking maintains spatial consistency despite active dynamic obstacles and severe stock depletion.

The full pipeline ran at 9.6Hz on a standard consumer laptop with a mid-tier GPU, achieving real-time throughput without LIDAR or wheel encoders.

#### F. Real-World Open-Vocabulary Evaluation

To demonstrate ShelfAware’s scalability and its independence from closed-set object detectors, we deployed the open-vocabulary pipeline (Pipeline B; Sec. III-B) using the wearable form factor in a 3,500 sq. ft. operational specialty grocery store, generating a highly tractable 380 MB semantic map (Fig. 2, right). Because deploying a motion capture system was infeasible in a public retail space, ground truth was established via optimized RTAB-Map SLAM trajectories from the mapping phase. From this, we sampled 10 random evaluation trajectories ( $\sim 40$  seconds each at a normal walking pace). To ensure a fair comparison, we conducted a grid search over filter parameters for all baselines, reporting the highest-performing configurations.

Table III presents these results. In this large, geometrically aliased environment, purely geometric methods (MCL, AMCL) collapse ( $\leq 8\%$  success). Using the semantic vision pipeline alongside depth observations, ShelfAware successfully achieved global localization in 62% of trials, significantly outperforming all baselines. We note that ShelfAware requires more time to achieve initial global convergence in this setting than the FQL baseline. This is an expected computational trade-off: to achieve its higher success rate across a 3,500 sq. ft. space, ShelfAware’s inverse semantic model must evaluate and inject particle proposals from a vastly larger candidate pose space.

We also performed an upper-bound ablation using ground-truth (GT) semantic vectors (obtained via raycasting the semantic map from the known RTAB-Map poses) while disabling the noisy depth likelihood. Under these perfect semantic conditions, ShelfAware achieved 98% global localization success and 92% tracking stability. In contrast, the FQL baseline only achieved 82% and 70% under the same perfect-vision conditions. This ablation proves that our core methodological contribution of modeling semantics as continuous probability distributions over object counts, inherently extracts more robust localization signals than fixed-quantity representations, even when perception is flawless. Furthermore, the ablation revealed that while depth observations are necessary to stabilize the noisy visual features in the “real” condition, incorporating depth actually degrades performance when semantic perception is perfect, highlighting the distinct aliasing risks of depth sensors in geometrically ambiguous domains.

#### G. Discussion and Limitations

*Implications for assistive devices for People with Visual Impairment (PVI):* Although our core contribution is a general method for vision-based localization in quasi-static environments, the results have direct implications for assistive navigation. First, *start-anytime* operation is crucial in shared-control assistive use: users often want to invoke assistance on demand rather than run a fully autonomous device continuously. ShelfAware’s rapid global localization (mean 1.07s across scenarios; Table I) and its ability to recover from lost tracking (Table II) align with these constraints by enabling on-demand pose estimation and re-localization without external infrastructure. Second, our chosen form factors reflect

TABLE I: Global localization results across setups in the mock store environment. Success is % of 25 trials per setup. The convergence time and RMSE is calculated only for successful convergences.

Method	Cart			Wearable			Dynamic			Degraded/Sparse			Overall G%
	Success	Time (s)	RMSE (m/rad)	Success	Time (s)	RMSE (m/rad)	Success	Time (s)	RMSE (m/rad)	Success	Time (s)	RMSE (m/rad)	
MCL	16%	1.06	0.41/0.27	12%	0.13	0.37/0.49	16%	0.13	0.25/0.04	16%	0.05	0.36/0.26	15%
AMCL	20%	17.41	<b>0.18/0.02</b>	0%	-	-	0%	-	-	20%	3.36	0.34/ <b>0.05</b>	10%
FQL	96%	1.98	0.38/0.20	96%	4.14	<b>0.42/0.30</b>	80%	2.75	<b>0.36/0.22</b>	88%	4.07	0.42/ <b>0.23</b>	90%
ShelfAware	<b>100%</b>	<b>0.27</b>	0.37/0.25	<b>100%</b>	<b>1.96</b>	0.36/0.42	<b>100%</b>	<b>0.29</b>	0.39/ <b>0.15</b>	<b>88%</b>	<b>1.83</b>	<b>0.31/0.34</b>	<b>97%</b>

TABLE II: Tracking Success Rate (T%) in the mock store.

Method	Cart	Wearable	Dynamic	Sparse	All	T-RMSE
MCL	8%	8%	12%	12%	10%	0.27/0.13
AMCL	17%	0%	0%	20%	10%	<b>0.25/0.14</b>
FQL	<b>60%</b>	60%	68%	48%	59%	0.32/0.10
<b>ShelfAware</b>	<b>60%</b>	<b>72%</b>	<b>80%</b>	<b>52%</b>	<b>66%</b>	<b>0.29/0.09</b>

TABLE III: Real-world grocery store evaluation. “Upper-Bound Semantics” isolates the filter using ground-truth semantics, while “Real Vision” uses the live RGB image. Metrics include Global Localization Success (G%), Tracking Success (T%), Time to Global Convergence (G-Time), and translational/rotational RMSE (m/rad).

Method	G%	T%	G-RMSE	T-RMSE	G-Time(s)
MCL	8%	6%	<b>0.30/0.25</b>	<b>0.31/0.15</b>	17.4
AMCL	6%	4%	0.36/0.13	0.41/ <b>0.03</b>	4.1
<i>Real Vision (RGB + Depth)</i>					
FQL	42%	38%	0.41/ <b>0.09</b>	0.37/0.06	<b>2.9</b>
<b>ShelfAware</b>	<b>62%</b>	<b>40%</b>	0.34/0.12	<b>0.31/0.10</b>	6.2
<i>Upper-Bound Semantics (GT Semantics, No Depth)</i>					
FQL	82%	70%	0.41/0.31	<b>0.14/0.07</b>	<b>1.6</b>
<b>ShelfAware</b>	<b>98%</b>	<b>92%</b>	<b>0.35/0.27</b>	0.19/ <b>0.06</b>	3.2

evidence on acceptance and usability: cane-mounted sensors face handling barriers and are often undesired by users [27] and head-mounted systems are frequently rejected for bulk and stigma [28], while cart-mounted or less bulky wearable solutions have shown promise [33]. These findings motivate the chest-mounted and cart configurations in Fig. 5 and Sec. III-F. Finally, a class-level semantic representation is a practical fit for retail navigation: detectors can see many *classes*, but mapping detections into a compact set is both stable under SKU churn and sufficiently distinctive for localization, as evidenced by the high success rates in Table I.

*Path to an assistive navigation stack:* ShelfAware provides spatial grounding that upstream guidance and downstream interaction modules can exploit. In shopping contexts, prior work has focused on product retrieval or fine-grained identification near the correct shelf [19], [33]. Our results address the prerequisite of reliably *reaching* the correct aisle/shelf in semantically dynamic, geometrically repetitive layouts. Integrating ShelfAware with wayfinding, obstacle avoidance, and product-retrieval interfaces (e.g., speech or haptics) is a natural next step toward end-to-end assistive experiences. Importantly, because our approach requires only vision sensors and VIO, it avoids environmental augmentation (e.g., RFID/beacons) and aligns with infrastructure-free

deployments [16]–[18], [33].

*Limitations:* Four limitations warrant discussion. (i) *Large-scale drift and map updates.* Our inverse proposal mechanism can overcome moderate semantic drift by re-injecting particles, provided sufficient semantic cues remain visible in the live scene. However, massive inventory reorganizations outstrip this capacity. ShelfAware currently lacks an automated update mechanism and significant changes require rebuilding the semantic map from scratch. (ii) *Perception failures.* Prolonged occlusions, motion blur, or severe category sparsity can reduce the information mass in the semantic vector, delaying inverse proposals and weakening forward likelihoods. (iii) *Computational complexity.* Precomputing the inverse semantic cache is strictly offline and scales linearly with the number of discrete states. Online, evaluating  $K$  depth beams and  $C$  semantic classes per particle scales linearly as  $\mathcal{O}(N)$ , since  $K$  and  $C$  are small constants. Recovering via particle injection takes only  $\mathcal{O}(1)$  hash-map lookups. (iv) *Human-factors validation.* Our form-factor choices are informed by prior literature and informal feedback, but we did not conduct PVI user studies; measuring usability and trust with PVI participants, including multi-hour battery tests and interaction design, is essential future work.

## V. CONCLUSION

In this work we propose *ShelfAware*, a novel method to address the challenging problem of vision-based localization in *quasi-static* indoor environments, where geometry is repetitive and local semantics evolve. ShelfAware models semantics as *distributional evidence over object categories* and couples this representation with an inverse semantic proposal mechanism inside an MCL framework, enabling the filter to remain informative under semantic drift and to generate targeted global pose hypotheses when needed.

Our core contributions are twofold: (i) a semantic mapping and observation design that encodes category counts and coarse spatial structure as probabilistic distributions; and (ii) a real-time particle filter that fuses depth likelihoods with this semantic similarity, utilizing precomputed semantic viewpoints for inverse pose proposals. We demonstrate the efficacy through extensive empirical evaluation, achieving robust *global localization and tracking* across both a highly controlled mock setup and a real-world grocery store.

In our mock environment, ShelfAware achieved **97%** overall global-localization success across four conditions vastly outperforming geometric baselines (15% MCL, 10% AMCL) and deterministic Fixed Quantity Landmark baselines (90% FQL). Furthermore, when deployed in a massive 3,500 sq.

ft. retail store using an open-vocabulary vision pipeline, it successfully globally localized in **62%** of trials (compared to FQL's 42%), proving its scalability and robustness to severe perception noise. An upper-bound ablation revealed that ShelfAware extracts significantly more localization signal from perfect semantic data than deterministic methods when provided with ground-truth semantics, ShelfAware achieved **98%** global success and **92%** tracking success, compared to just **82%** and **70%** for the FQL.

Beyond retail, the method applies to service and mobile robots in warehouses and offices, where quasi-static semantics and ambiguous geometry are common. In assistive contexts, the ability to *start anytime* and *relocalize* quickly supports shared-control and infrastructure-free deployment. Future work includes larger-scale evaluations across diverse layouts, embedded implementations, and online maintenance of the semantic layer, further advancing practical, vision-only localization in dynamic real-world environments.

## REFERENCES

- [1] D. Fox, "KLD-Sampling: Adaptive Particle Filters," in *Advances in Neural Information Processing Systems*, vol. 14. MIT Press, 2001.
- [2] C. P. Gharpure and V. A. Kulyukin, "Robot-assisted shopping for the blind: issues in spatial cognition and product selection," *Intelligent Service Robotics*, vol. 1, no. 3, pp. 237–251, July 2008.
- [3] N. Zimmerman, L. Wiesmann, T. Guadagnino, T. Labe, J. Behley, and C. Stachniss, "Robust Onboard Localization in Changing Environments Exploiting Text Spotting," July 2022, arXiv:2203.12647 [cs].
- [4] N. Zimmerman, T. Guadagnino, X. Chen, J. Behley, and C. Stachniss, "Long-Term Localization Using Semantic Cues in Floor Plan Maps," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 176–183, Jan. 2023, conference Name: IEEE Robotics and Automation Letters.
- [5] R. G. Goswami, P. V. Amith, J. Hari, A. Dhaygude, P. Krishnamurthy, J. Rizzo, A. Tzes, and F. Khorrani, "Efficient Real-Time Localization in Prior Indoor Maps Using Semantic SLAM," in *2023 9th International Conference on Automation, Robotics and Applications (ICARA)*, Feb. 2023, pp. 299–303, iSSN: 2767-7745.
- [6] F. Xie and S. Schwertfeger, "Robust Lifelong Indoor LiDAR Localization Using the Area Graph," *IEEE Robotics and Automation Letters*, Jan. 2024, conference Name: IEEE Robotics and Automation Letters.
- [7] "AMCL nav2 documentation," <https://docs.nav2.org/configuration/packages/configuring-amcl.html>, accessed 2025-09-21.
- [8] 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," *International Journal of Computer Vision*, vol. 132, no. 8, pp. 3139–3171, Aug. 2024.
- [9] H. Kuang, X. Chen, T. Guadagnino, N. Zimmerman, J. Behley, and C. Stachniss, "IR-MCL: Implicit Representation-Based Online Global Localization," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, Mar. 2023, conference Name: IEEE Robotics and Automation Letters.
- [10] R. Crabb, S. A. Cheraghi, and J. M. Coughlan, "A Lightweight Approach to Localization for Blind and Visually Impaired Travelers," *Sensors*, vol. 23, no. 5, p. 2701, Jan. 2023.
- [11] C. Rui, Y. Liu, J. Shen, Z. Li, and Z. Xie, "A Multi-Sensory Blind Guidance System Based on YOLO and ORB-SLAM," in *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)*, Dec. 2021, pp. 409–414, iSSN: 2329-6259.
- [12] H. Chen, Y. Zhang, K. Yang, M. Martinez, K. Muller, and R. Stiefelhagen, "Can We Unify Perception and Localization in Assisted Navigation? An Indoor Semantic Visual Positioning System for Visually Impaired People," in *Computers Helping People with Special Needs*. Cham: Springer International Publishing, 2020.
- [13] T. Loiseau, G. Bourmaud, and V. Lepetit, "Alligat0r: Pre-training through co-visibility segmentation for relative camera pose regression," *arXiv preprint arXiv:2503.07561*, 2025.
- [14] S. Dong, S. Wang, S. Liu, L. Cai, Q. Fan, J. Kannala, and Y. Yang, "Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
- [15] B. Al-Tawil, T. Hempel, A. Abdelrahman, and A. Al-Hamadi, "A review of visual slam for robotics: Evolution, properties, and future applications," *Frontiers in Robotics and AI*, vol. 11, p. 1347985, 2024.
- [16] V. Kulyukin, C. Gharpure, and J. Nicholson, "RoboCart: toward robot-assisted navigation of grocery stores by the visually impaired," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug. 2005, pp. 2845–2850, iSSN: 2153-0866.
- [17] D. Lopez-de Ipia, T. Lorigo, and U. Lopez, "BlindShopping: Enabling Accessible Shopping for Visually Impaired People through Mobile Technologies," in *Toward Useful Services for Elderly and People with Disabilities*, B. Abdulrazak, S. Giroux, B. Bouchard, H. Pigot, and M. Mokhtari, Eds. Berlin, Heidelberg: Springer, 2011.
- [18] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh, "Vizwiz: Locateit-enabling blind people to locate objects in their environment," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 65–72.
- [19] S. Agrawal, S. Nayak, A. Naik, and B. Hayes, "ShelfHelp: Empowering Humans to Perform Vision-Independent Manipulation Tasks with a Socially Assistive Robotic Cane," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, May 2023, pp. 1514–1523.
- [20] R. Kamikubo, H. Kacorri, and C. Asakawa, "'we are at the mercy of others' opinion": Supporting blind people in recreational window shopping with ai-infused technology," in *Proceedings of the 21st International Web for All Conference*, 2024, pp. 45–58.
- [21] Y. Kaniwa, M. Kuribayashi, S. Kayukawa, D. Sato, H. Takagi, C. Asakawa, and S. Morishima, "Chitchatguide: How can a guidance system with large language models impact shopping mall experiences for people with visual impairments?" in *International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2024.
- [22] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.
- [23] E. Goldman, R. Herzig, A. Eisenschtag, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *Proceedings of the CVF conference on computer vision and pattern recognition*, 2019.
- [24] P. Koopman, "Bresenham line-drawing algorithm," *Forth Dimensions*, vol. 8, no. 6, pp. 12–16, 1987.
- [25] C. H. Walsh and S. Karaman, "Cddt: Fast approximate 2d ray casting for accelerated localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3677–3684.
- [26] X. Liu, Z. Zhang, J. Peterson, and S. Chandra, "The effect of lidar data density on dem accuracy," in *Proceedings of the 17th International Congress on Modelling and Simulation (MODSIM07)*. Modelling and Simulation Society of Australia and New Zealand, 2007.
- [27] S. Y. Kim and K. Cho, "Usability and design guidelines of smart canes for users with visual impairments," *international Journal of Design*, vol. 7, no. 1, 2013.
- [28] K. Lee, D. Sato, S. Asakawa, H. Kacorri, and C. Asakawa, "Pedestrian detection with wearable cameras for the blind: A two-way perspective," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [29] Y. Zhang, Z. Li, H. Guo, L. Wang, Q. Chen, W. Jiang, M. Fan, G. Zhou, and J. Gong, "'i am the follower, also the boss': Exploring different levels of autonomy and machine forms of guiding robots for the visually impaired," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–22.
- [30] R. Kamikubo, S. Kayukawa, Y. Kaniwa, A. Wang, H. Kacorri, H. Takagi, and C. Asakawa, "Beyond omakase: Designing shared control for navigation robots with blind people," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.
- [31] A. Arora, L. Nadolskis, M. Beyeler, and M. Sra, "Visionai-shopping assistance for people with vision impairments," in *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2024, pp. 377–378.
- [32] K. G. Derpanis, "Overview of the ransac algorithm," *Image Rochester NY*, vol. 4, no. 1, pp. 2–3, 2010.
- [33] J. Nicholson, V. Kulyukin, and D. Coster, "Shoptalk: independent blind shopping through verbal route directions and barcode scans," *The Open Rehabilitation Journal*, vol. 2, no. 1, pp. 11–23, 2009.
- [34] A. K. Sengupta and B. Das, "Maximum reach envelope for the seated and standing male and female for industrial workstation design," *Ergonomics*, vol. 43, no. 9, pp. 1390–1404, 2000.