

Anticipating Human Behavior for Safe Navigation and Efficient Collaborative Manipulation with Mobile Service Robots

Simon Bultmann, Raphael Memmesheimer, Jan Nogga, Julian Hau, and Sven Behnke

Abstract—The anticipation of human behavior is a crucial capability for robots to interact with humans safely and efficiently. We employ a smart edge sensor network to provide global observations, future predictions, and goal information to integrate anticipatory behavior for the control of a mobile manipulation robot. We present approaches to anticipate human behavior in the context of safe navigation and collaborative mobile manipulation. First, we anticipate human motion by employing projections of predicted human trajectories from smart edge sensor observations into the planning map of a mobile robot. Second, we anticipate human intentions in a collaborative furniture-carrying task to achieve a given room layout. Our experiments indicate that anticipating human behavior allows for safer navigation and more efficient collaboration. Finally, we showcase an integrated robotic system that anticipates human behavior while collaborating with an operator to achieve a target room layout, including the placement of tables and chairs.

I. INTRODUCTION

Humans anticipate the actions of others in their surroundings and plan their own actions accordingly. This ability makes interaction more intuitive, efficient, safe, and natural. Robots, however, often lack this ability and are perceived as unpredictable, leading to unsafe interactions [1], [2].

While much research has been conducted on anticipating human behavior [1], [3]–[5], we found that the actual integration of anticipatory behavior into mobile manipulation robot planning remains underrepresented. This work presents approaches for anticipating human behavior in two natural interaction and collaboration scenarios.

First, we anticipate human behavior by encoding future human trajectories observed by an allocentric smart edge sensor network into the planning map of an autonomously navigating robot. This allows the robot to incorporate globally observed predictions into its local planning system and navigate more safely among humans. The smart edge sensor system enables the incorporation of predictions that cannot be anticipated with local sensory inputs alone.

Second, we anticipate human behavior in the context of collaborative furniture handling. In this scenario, a mobile manipulation platform collaborates with a human to move tables and chairs to achieve a predefined room layout. Anticipation in this context combines compliant collaborative control with goal anticipation given the target layout.

Prior existing work considered robots as nodes in a smart edge sensor network [6] for collaborative semantic mapping.

This work was funded by grant BE 2556/16-2 (Research Unit FOR 2535 Anticipating Human Behavior) of the German Research Foundation (DFG)

All authors are with the Autonomous Intelligent Systems group, University of Bonn, Germany; {bultmann|memmesheimer}@ais.uni-bonn.de

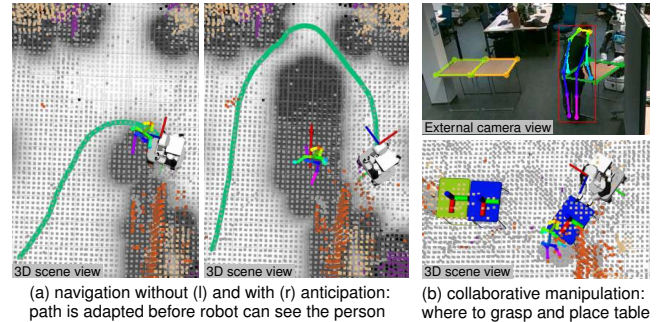


Fig. 1. Two scenarios in which our robot anticipates human behavior.

We build on this, but now focus on anticipating human behavior based on the allocentric 3D semantic scene model built from smart edge sensor observations and making the robot act on these anticipations.

Fig. 1 illustrates the two scenarios in which we anticipate human behavior. For the navigation scenario, we anticipate future human trajectories and adapt the robot's navigation path accordingly. For the collaborative manipulation scenario, we anticipate the human's intention to decide from where to grasp which table. During carrying, we incorporate the anticipated target pose for collaborative manipulation.

The main contributions of this paper are as follows:

- We propose an approach for anticipating human motion for safe, human-aware navigation.
- We present an approach for anticipating collaborative human behavior in the context of furniture carrying to achieve a given target configuration.
- We conduct real-robot experiments with two subjects to evaluate the performance of our proposed methods.
- We successfully demonstrate the benefits of anticipating human behavior in the context of safe navigation and collaborative manipulation with a mobile service robot.

II. RELATED WORK

◦ *Anticipation*: Canuto *et al.* [4] analyze the influence of contextual cues (gaze, movement, object information) to improve action anticipation in a collaborative setting. The problem is formulated as a classification task to predict the next most likely action. Duarte *et al.* [5] conducted experiments within a similar setup where humans had to anticipate the actions of a robot. Tanke *et al.* [7] predict future human movements using a diffusion model where the loss function emphasizes the importance of social context. In contrast to the above approaches, which primarily analyze anticipation as a prediction problem, there is relatively little work on integrating anticipation into robotic systems, where robots

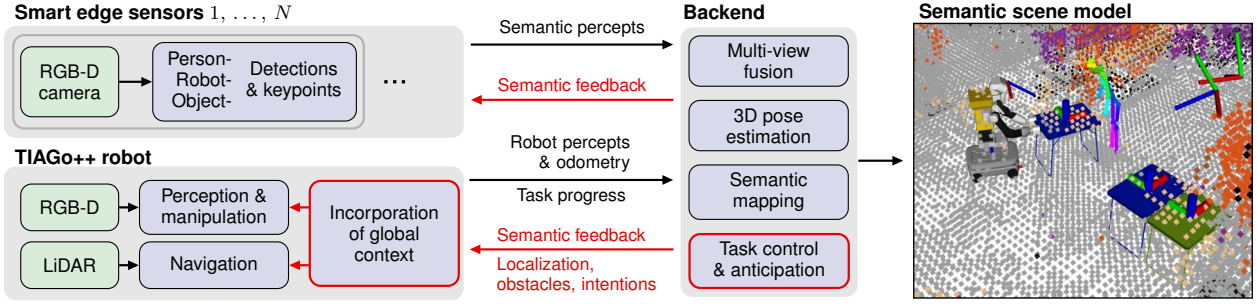


Fig. 2. Data processing architecture of the developed approach. A network of smart edge sensors supervises the work space. It detects persons, robots, and objects and estimates their pose. The backend fuses local percepts and controls the task. Both the robot and the sensors incorporate semantic feedback.

adapt their behaviour based on predicted human intentions. Huang and Mutlu [2] present a system that proactively performs actions to assist a human in a collaborative task based on the user’s gaze patterns. Like our approaches, they observed an improved task completion time. Psarakis *et al.* [3] study human-robot collaboration in an industrial setting, focusing on the effects of robot anticipatory cues and adaptability on task efficiency, safety, and fluency. Bärmann *et al.* [8] present an approach for dialog-based incremental learning through natural interaction. Large language models generate high-level behavior that adapts to future tasks.

◦ *Human-Aware Navigation:* de Heuvel *et al.* [9] adapt robot navigation to user preferences that have been explicitly demonstrated in a virtual-reality environment beforehand. The approach is evaluated in simulation only. Luber *et al.* [10] present a learning-based approach for social-aware robot navigation. Criteria like path length and travel time, as well as objective criteria such as social comfort, are addressed by an unsupervised learning approach. Bruckschen *et al.* [11] present a foresighted navigation approach, ensuring social comfort while anticipating human trajectories at a target location to offer timely assistance. In contrast to our approach, they rely on the onboard sensors of the robot, while our approach integrates predictions by a smart edge sensor network that allows for incorporating predictions that are not tractable by relying solely on the robot’s observations. Other approaches employ social force models to navigate among humans [12], [13]. Arena *et al.* [14] learn to navigate and anticipate actions in an environment based on sensory input. The anticipation of human behavior is not considered.

◦ *Furniture Handling:* Rus *et al.* [15] were among the first to present an approach for collaborative mobile furniture movement. In their setting, in modern terms, simplistic robots were moving furniture by executing push sequences and aiming to enforce contact with furniture. Knepper *et al.* [16] presented an approach for coordinated multi-robot furniture assembly. A geometric specification is provided to their approach to generate a symbolic plan for the assembly of parts. The assembly plan is then sequentially executed by two collaborating robots equipped with special tool end-effectors. The vision part is bypassed by using motion capture systems. The approach has been exemplarily demonstrated for the assembly of an Ikea Lack table. A quantitative evaluation of the system has not been conducted. Closely related to our

proposed approach is the work of Stückler and Behnke [17] and Stückler *et al.* [18]. They presented a compliant system for various manipulation tasks, including compliance control of human collaborative table carrying. Upon detecting a person, the robot proceeds to approach and grasp the table, subsequently waiting for the individual to initiate lifting. When the person lowers the table, the robot initiates a placement motion. In contrast to our method, the robot does not actively anticipate the intended outcome. The evaluation of the system is limited to a qualitative demonstration of the system in the context of the RoboCup@Home competition. Fallatah *et al.* [19] present an approach for furniture arrangement of chairs. A multipurpose screen setup is used to set the arrangement. An overhead camera is employed to visualize the scenes on the screen and provides feedback to the robot. The chairs are attached directly to the robots, facilitating their manipulation, and markers are attached to the chairs, simplifying the perception. In contrast to the previous approaches, a user study was conducted to evaluate whether user expectations were met.

III. METHOD

A. Overview

An overview of the data processing architecture of our proposed approach is given in Fig. 2. A network of 25 smart edge sensors (cf. Fig. 6 (b)) is installed in our lab and gathers semantic observations of the scene, i.e., detections and keypoints of persons, robots, and objects, as well as semantic point clouds [20]–[22]. The sensor views are fused on a central backend into an allocentric 3D semantic scene model \mathcal{M} comprising a volumetric semantic map \mathcal{V} of the static environment as well as dynamic human, robot, and object models $\mathcal{P}, \mathcal{R}, \mathcal{O}$, where:

$$\mathcal{P} = \left\{ \left(\{ \mathbf{kp} \}_{i=1}^K \in \mathbb{R}^3, \mathbf{v} \in \mathbb{R}^3 \right) \right\}, \quad (1)$$

$$\mathcal{R} = \left(\mathbf{M}_r, \mathbf{p}_r = [x_r, y_r, \theta_r]^\top \in \text{SE}(2) \right), \quad (2)$$

$$\mathbf{v}_r = [v_x, v_y, \omega_\theta]^\top \in \mathbb{R}^3),$$

$$\mathcal{O} = \left\{ \left(\mathbf{M}_o, c, \mathbf{p}_o = [x_o, y_o, \theta_o]^\top \in \text{SE}(2) \right) \right\}. \quad (3)$$

The tracked persons \mathcal{P} are represented by a set of K 3D keypoints $\{ \mathbf{kp} \}_{i=1}^K$ that form the body skeleton and the root joint velocity \mathbf{v} . The robot \mathcal{R} is represented by its mesh

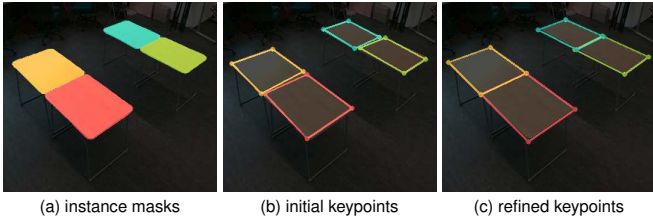


Fig. 3. Table instance segmentation from smart edge sensor view: extracted masks (a) and keypoints (b) before, and (c) after edge alignment.

model M_r , pose \mathbf{p}_r , and velocity \mathbf{v}_r on the ground plane. We assume a single robot visible in the scene. The objects \mathcal{O} are represented by their mesh model M_o , semantic class c , and pose \mathbf{p}_o on the ground plane. We consider tables and chairs as object classes in our model.

The smart edge sensors receive parts of the semantic scene model \mathcal{M} as semantic feedback to incorporate global context, e.g., about occlusions, into their local perception [20], [21]. The robot, similar to the smart edge sensors, augments its local perception, manipulation, and navigation capabilities with global context information received as semantic feedback from the semantic scene model \mathcal{M} .

The robot receives feedback about (i) its pose in the scene model, (ii) persons who are in its vicinity but are out of sight of its internal sensors, e.g., due to occlusions or limited Field of View (FoV), and their predicted movement, as well as (iii) the objects and the intended target configuration of the manipulation task. This enables anticipatory human-aware robot navigation where the robot preemptively adjusts its navigation path, e.g., to persons appearing from behind occluders (cf. Algorithm 1) or to reach the intended target pose for picking up or placing an object (cf. Algorithms 2, 3).

We detail the developed approaches for perception, navigation, and manipulation in the following.

B. Semantic Perception

◦ *Smart Edge Sensor Network*: The smart edge sensor boards gather semantic observations from multiple views, covering the entire scene from different directions. Thus, a complete, allocentric 3D semantic scene model \mathcal{M} is estimated and updated over time, based on [20]–[22], comprising a semantic map of the static environment \mathcal{V} as well as dynamic human, robot, and object poses \mathcal{P} , \mathcal{R} , \mathcal{O} (Eqns. (1)–(3)).

For robot localization by the external smart edge sensors, we adapt prior work [6], developed for the Toyota HSR robot, to the employed TIAGo++ robot. While in [6], only static keypoints on the rigid body were used for robot pose estimation, here, we also employ dynamic keypoints on the articulated robot body and dynamically adapt the reference points for pose estimation based on the robot joint configuration. The pose \mathbf{p}_r from the robot model \mathcal{R} (Eq. (2)) estimated by the sensor network is sent to the robot to initialize and correct its localization in the scene.

We further extend the semantic perception to furniture objects (i.e., tables, chairs), with a focus on separating object instances that stand closely together side-by-side.

To separate individual objects in the local sensor views, we employ an instance segmentation model based on YOLO-

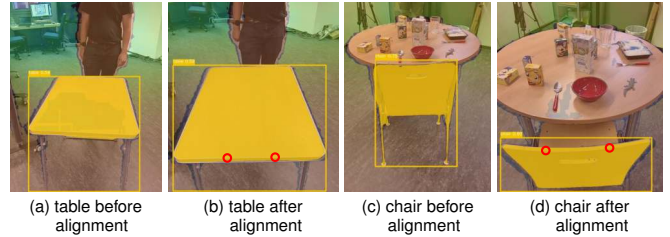


Fig. 4. Local semantic perception, before and after alignment to table and chair. The red circles denote the computed grasping points.

v8 [23], fine-tuned on synthetic data generated with Omniverse Replicator [24]. For the table class, the model is trained to segment only the tabletop, allowing keypoints to be directly extracted from the instance masks (Fig. 3 (a)). Initial table keypoints are obtained from the mask contours with the Douglas-Peucker algorithm [25], selecting four points that maximize the Intersection over Union (IoU) with the mask (Fig. 3 (b)). The resulting quadrilateral is refined by aligning its edges with the mask contour to precisely determine the four corner keypoints of each tabletop (Fig. 3 (c)). Object poses are estimated from 2D-3D keypoint correspondences via PnP [26], [27], restricted to 3 DoF on the ground plane. The table yaw angle θ_t is further restrained to $[0 \dots \pi]$ to account for their 180° symmetry. After removing outlier pose estimates, ICP refinement is performed between the mesh and the instance point cloud. Chair keypoints and poses are computed following [22], using chair crops derived from the instance segmentation. The resulting object poses are transmitted to the central backend, where they are fused with other sensor views and tracked using a Kalman filter [28].

The obtained object models \mathcal{O} (Eq. (3)) are used together with the tracked persons \mathcal{P} (Eq. (1)) [20] to anticipate the robot target poses for pickup and placement of the furniture objects (cf. Sec. III-C).

◦ *Robot*: To plan the robot’s arm motions to pick up the targeted furniture objects, the robot uses local, onboard semantic perception from its RGB-D camera. Upon reaching the goal pose received from the sensor network (cf. Sec. III-C), the open vocabulary detector MM-Grounding-Dino [29] is prompted in a zero-shot manner for tables or chairs in the RGB image. The resulting object bounding boxes are forwarded to Nano SAM [30], yielding a segmentation mask per object. Since Nano SAM is a distilled, CNN-based version of Mobile SAM [31], a lightweight version of SAM [32], this step adds little computational overhead to the local perception pipeline. Individual object point clouds are obtained from the depth image using the segmentation masks. At this point, the closest object is selected for manipulation. For tables, the tabletop plane is extracted and the corners are sorted to determine the line describing the front side of the table. Grasp poses are determined to symmetrically position the grippers parallel around the center of the table front. For chairs, the backrest is extracted from the point cloud and a line is fitted to its cross-section in the horizontal plane. Grasp poses are determined symmetrically around the center of this line and projected into the backrest cross-section.

Algorithm 1 Anticipatory Navigation Cost Map

Input: $t_{\text{pred}}, t_{\text{step}}$ # prediction time horizon and step
 $\mathcal{P} = \{(\{\mathbf{kp}\}, \mathbf{v})\}, \mathbf{p}_r$ # tracked persons; robot pose

- 1: $\{\mathbf{kp}_{\text{pred}}\} \leftarrow \emptyset$
- 2: $\mathcal{P}_{\text{valid}} \leftarrow \text{filter_outlier}(\mathcal{P})$ # remove far away or incomplete person tracks.
- 3: **for all** $P \in \mathcal{P}_{\text{valid}}$ **do**
- 4: $(\{\mathbf{kp}_{\text{local}}\}, \mathbf{v}_{\text{local}}) \leftarrow \text{transform_frame}(P, \mathbf{p}_r)$
- 5: $\{\mathbf{kp}_{\text{pred}}\} \leftarrow \{\mathbf{kp}_{\text{pred}}\} \cup \{\mathbf{kp}_{\text{local}}\}$
- 6: **for** $\Delta t = 0, t_{\text{step}}, \dots, t_{\text{pred}}$ **do**
- 7: $\{\mathbf{kp}_{\text{pred}}\} \leftarrow \{\mathbf{kp}_{\text{pred}}\} \cup \{\mathbf{kp}_{\text{local}} + \Delta t \cdot \mathbf{v}_{\text{local}}\}$
- 8: $\{\mathbf{kp}_{\text{out}}\} \leftarrow \text{inflate_and_proj_to_ground}(\{\mathbf{kp}_{\text{pred}}\})$
- 9: $\text{publish_to_robot}(\{\mathbf{kp}_{\text{out}}\})$ # publish point cloud used as additional input to navigation cost map.

Before executing the grasp, the robot base is moved to improve its alignment w.r.t. the estimated object orientation and distance, and the perception is repeated. Fig. 4 shows the local perception of the table and chair before and after alignment of the robot base to the objects.

C. Anticipatory Navigation

To realize anticipatory human-aware robot navigation, we integrate semantic feedback about persons and their velocities \mathcal{P} , tracked by the smart edge sensor network, into the local dynamic obstacle cost map of the robot navigation stack. This implementation on a low planning level generalizes well between different robots, and we integrate it into both HSR and TIAGo robots (cf. Sec. IV-B.1).

The local cost map is a 2D grid map in robot coordinates that incorporates multiple sensor sources (i.e., robot-internal 2D LiDARs and RGB-D camera) that provide information about occupied areas near the robot. It is used to dynamically update the navigation path to avoid obstacles. We add to it a virtual point cloud measurement as an additional sensor source, received as semantic feedback from the sensor network. This virtual point cloud comprises areas predicted to be occupied by the tracked persons.

The virtual point cloud calculation is detailed in Alg. 1: We employ 3D person keypoints $\{\mathbf{kp}\}_{i=1}^K$ and their resp. root joint velocity \mathbf{v} tracked by the sensor network. Person joints in the robot’s vicinity are transformed to robot-centric coordinates using the tracked robot pose \mathbf{p}_r . The transformed joints are extrapolated using \mathbf{v} by $t_{\text{pred}} = 2\text{s}$, empirically balancing prediction horizon and cost map compactness for typical walking speeds. Finally, they are inflated by a safety margin and projected onto the ground plane. The resulting point cloud $\{\mathbf{kp}_{\text{out}}\}$, representing regions currently or soon occupied by persons, is sent to the robot for integration into its dynamic obstacle avoidance cost map.

This enables anticipatory human-aware robot navigation, where the robot foresightedly adapts its path, e.g., to persons emerging from occlusions. Fusing onboard sensing with semantic feedback from the instrumented environment allows the robot to effectively “see around corners”.

For collaborative furniture handling, we anticipate the human’s intention for where to pick up and place the furniture

Algorithm 2 Pickup Pose Anticipation

Input: $\mathcal{P} = \{\mathbf{x}_p\}$ # tracked persons (root kps only)
 $\mathcal{T} = \{[\mathbf{x}_t, \theta_t]^\top\}, \mathcal{A}_{\text{pick}}$ # tracked tables; pickup area
 $\mathbf{p}_r, \mathbf{v}_r, \delta_{\text{grasp}}$ # robot pose & velocity; grasp distance
 $\tau_p, \tau_{\text{goal}}, \tau_{\text{vel}}$ # pickup and goal pos & vel. thresholds

- 1: **while** True **do**
- 2: $\mathcal{T}_{\text{pick}} \leftarrow \text{select_tables}(\mathcal{T}, \mathcal{A}_{\text{pick}})$ # tables in pick area.
- 3: $P \leftarrow \text{select_person}(\mathcal{P}, \mathcal{A}_{\text{pick}}, \tau_p)$ # person closest to pickup area center and distance within τ_p .
- 4: **if** P not None **then**
- 5: $T \leftarrow \text{closest_table}(\mathcal{T}_{\text{pick}}, P)$
- 6: # Calculate goal pose to grasp table from opposite side than person.
- 7: $\mathbf{n}_t \leftarrow [\cos(\theta_t), \sin(\theta_t)]^\top$ # table long side.
- 8: $\mathbf{n}_p \leftarrow \mathbf{x}_t - \mathbf{x}_p$ # person approach vector.
- 9: $\mathbf{x}_{\text{goal}} \leftarrow \mathbf{x}_t + \text{sign}(\mathbf{n}_p^\top \mathbf{n}_t) \delta_{\text{grasp}} \mathbf{n}_t$
- 10: $\mathbf{p}_{\text{goal}} \leftarrow [\mathbf{x}_{\text{goal}}, \text{adjust_ori}(\theta_t, \text{sign}(\mathbf{n}_p^\top \mathbf{n}_t))]$
- 11: $\text{publish_to_robot}(\mathbf{p}_{\text{goal}})$ # send via network.
- 12: **if** $\text{pos_diff}(\mathbf{p}_{\text{goal}}, \mathbf{p}_r) < \tau_{\text{goal}}$ and $|\mathbf{v}_r| < \tau_{\text{vel}}$ **then**
- 13: **return**

based on a pre-defined pickup area $\mathcal{A}_{\text{pick}}$ and target layout \mathcal{L} . The pickup pose anticipation is detailed in Algorithm 2. When the sensor network detects a person P moving towards $\mathcal{A}_{\text{pick}}$, the robot receives a goal pose $\mathbf{p}_{\text{goal}} \in \mathbb{SE}(2)$ to navigate to the table T closest to the person. This goal pose is dynamically adjusted depending on the table side the person moves to, or if they move on to another table. The goal position \mathbf{x}_{goal} is offset from the table center position \mathbf{x}_t by the grasp distance δ_{grasp} along the table’s x-axis \mathbf{n}_t towards the opposite side the person is approaching from. The orientation is adjusted so that the robot faces the table. This ensures the robot aligns correctly for collaborative grasping. When the goal pose is reached, the robot base alignment is refined based on the robot’s onboard perception and the object is grasped as detailed in Sec. III-B and III-D.

D. Collaborative Manipulation

Given a viable alignment of the robot base, a bimanual grasping motion is performed by bringing the robot arms forward from their home configuration and moving the end-effectors to pre-grasp orientations for tables or chairs using predefined joint space commands. Next, a two-step Cartesian trajectory is computed and executed based on the grasp poses calculated by the local semantic perception. This trajectory moves the end-effectors to their grasp poses while ensuring that the final approach is exclusively along the open direction of the grippers to prevent collisions with the object surface.

Once the table is grasped, the robot anticipates the human’s intention to lift it and activates a compliant control mode by switching its arms to a gravity-compensation mode, configured to hold the weight of the table. The person on the other side of the table can move the robot’s arms through it to enable collaborative carrying of the furniture.

Velocity commands $\mathbf{v}_{\text{ee}} = [v_x, v_y, \omega_\theta]^\top$ for both linear and angular robot base movements are calculated proportionally to the end-effector displacement $\Delta \mathbf{x}_{\text{ee}} = [\Delta x_{\text{ee}}, \Delta y_{\text{ee}}]^\top$

Algorithm 3 Collaborative Carrying and Goal Anticipation

Input: $\mathbf{x}_{ee,init}, \mathbf{x}_{ee}$ # *init. & curr. end-effector position*
 $\mathcal{L} = \{[\mathbf{x}_t, \theta_t]^\top\}, \mathbf{p}_r = [\mathbf{x}_r, \theta_r]^\top$ # *layout; robot pose*
 $\tau_{ee}, \tau_{goal}, \tau_{direct}$ # *thresholds*
 $k_{e,lin}, k_{e,rot}, k_{a,lin}, k_{a,rot}, k_{direct}$ # *scaling factors*

```

1: while True do
2:    $\mathbf{p}_{goal} \leftarrow \text{update\_goal}(\mathcal{L}, \mathbf{p}_r)$  # robot pose to place table at layout pose closest to current position.
3:    $\text{look\_at}(\mathbf{p}_{goal})$  # robot head looks at goal.
4:    $\Delta \mathbf{x}_{ee} \leftarrow \mathbf{x}_{ee} - \mathbf{x}_{ee,init}$  # end-effector displacement.
5:    $\Delta \mathbf{x}_{goal} \leftarrow \mathbf{x}_{goal} - \mathbf{x}_r$  # vector to goal.
6:    $\Delta \theta_{goal} \leftarrow \text{angle\_diff}(\theta_{goal} - \theta_r)$ 
7:    $\mathbf{v}_{ee} \leftarrow \mathbf{0}, \mathbf{v}_a \leftarrow \mathbf{0}, \mathbf{v}_r \leftarrow \mathbf{0}$  #  $\mathbf{v} = [v_x, v_y, \omega_\theta]^\top$ .
8:   if  $\|\Delta \mathbf{x}_{goal}\| < \tau_{goal}$  then # goal reached.
9:      $\text{publish}(\mathbf{v}_r)$ 
10:    return
11:   if  $|\Delta x_{ee}| > \tau_{ee}$  then # forward movement.
12:      $v_{ee,x} \leftarrow k_{e,lin} \cdot \Delta x_{ee}$ 
13:   if  $|\Delta y_{ee}| > \tau_{ee}$  then # lateral movement or rotation.
14:     if  $|\Delta x_{ee}| > \tau_{ee}$  then
15:        $\omega_{ee,\theta} \leftarrow k_{e,rot} \cdot \Delta y_{ee}$  # when moving forward, use lateral  $\Delta y_{ee}$  for turning.
16:     else # else, use lateral  $\Delta y_{ee}$  for sideways motion.
17:        $v_{ee,y} \leftarrow k_{e,lin} \cdot \Delta y_{ee}$ 
18:    $\mathbf{v}_{a,xy} \leftarrow k_{a,lin} \cdot \Delta \mathbf{x}_{goal}$  # linear velocity to goal.
19:    $\omega_{a,\theta} \leftarrow k_{a,rot} \cdot \Delta \theta_{goal}$  # angular velocity to goal.
20:   if  $\|\Delta \mathbf{x}_{goal}\| > \tau_{direct}$  or  $|\Delta x_{ee}| > \tau_{ee}$  then
21:      $\mathbf{v}_r \leftarrow \mathbf{v}_{ee} + \mathbf{v}_a$  # far from goal or moving: combine end-effector and goal velocities
22:   else # close to goal and stopped: direct approach.
23:      $\mathbf{v}_r \leftarrow k_{direct} \cdot \mathbf{v}_a$  # scale up vel. due to small diffs.
24:      $\text{publish}(\mathbf{v}_r)$ 

```

as detailed in Algorithm 3. A forward displacement Δx_{ee} larger than τ_{ee} will induce forward motion, and a sideways displacement Δy_{ee} larger than τ_{ee} a turn when moving forward, or a sideways movement when there is no forward movement (Alg. 3, ll. 11-17). This enables the person to control all degrees of freedom of the omnidirectional robot base for guiding the robot toward the target pose.

The robot further anticipates the target pose to place the table from the furniture layout \mathcal{L} . The robot head turns to look at the selected goal to signal the anticipation to the operator. The velocity \mathbf{v}_{ee} is overlaid with a velocity vector \mathbf{v}_a pointing towards the anticipated goal (Alg. 3, ll. 18-21). Thus, the robot automatically draws towards the placement location but can still be influenced by the person, e.g., to avoid obstacles or towards a different target pose from \mathcal{L} . Fig. 5 illustrates the collaborative carrying.

Once the target comes closer than a threshold $\tau_{direct} = 1$ m and the operator stops the robot by moving the end-effectors back to the initial position $\mathbf{x}_{ee,init}$, the robot moves the last part fully automatically (Alg. 3 ll. 22-23), to achieve a precise positioning of the furniture. When the target pose is reached within tolerance τ_{goal} , the robot stops the compliant mode and executes the placement motion, which is the reverse of

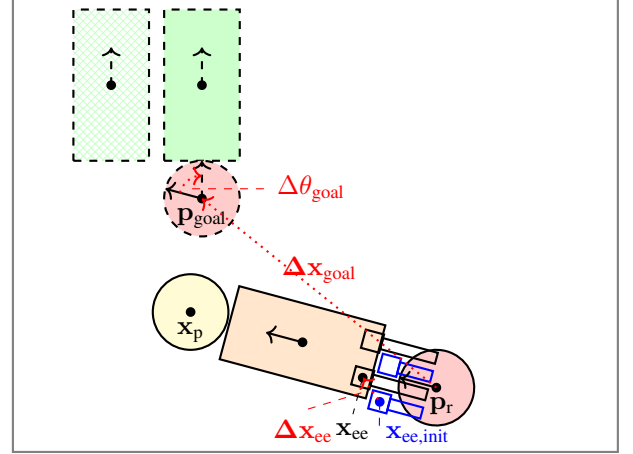


Fig. 5. Collaborative carrying with goal anticipation including robot pose \mathbf{p}_r , human position \mathbf{x}_p , the target table layout (green), anticipated goal pose \mathbf{p}_{goal} , and the end-effector initial and current positions $\mathbf{x}_{ee,init}$ resp. \mathbf{x}_{ee} . The robot velocity is determined w.r.t. the position and orientation difference to the goal $\Delta \mathbf{x}_{goal}$ resp. $\Delta \theta_{goal}$, and the end-effector displacement $\Delta \mathbf{x}_{ee}$.

the grasp motion. After placement, the now occupied target pose is removed from the layout plan \mathcal{L} , and the robot moves on to pick up the next object.

The chairs, in contrast to the tables, are handled by the robot alone, without compliant movement control, directly pulling and pushing them from pickup to placement location.

IV. EXPERIMENTAL EVALUATION

A. Setup

For our evaluation experiments, we use a PAL Robotics TIAGo++ [33] omnidirectional dual-arm mobile manipulator, shown in Fig. 6(a), equipped with an Orbbeo Gemini 335 RGB-D camera and a Zotac ZBOX QTG7A4500 computer mounted on the back of the robot.

The robot is informed by a smart edge sensor network consisting of 25 sensor nodes, shown in Fig. 6(b), with an Intel RealSense D455 RGB-D camera and an Nvidia Jetson Orin Nano compute board with an embedded GPU for onboard semantic perception using lightweight CNNs [21], [22]. The smart edge sensors are mounted at a height of ~ 2.5 m, distributed over a lab space of ~ 240 m² size.

B. Evaluation

Many approaches related to furniture manipulation focus on qualitative demonstrations. In this work, we conduct repetitive experiments following a task protocol for both anticipatory navigation and collaborative furniture arrangement.

1) *Human-Aware Anticipatory Navigation:* In our anticipatory human-aware robot navigation experiment, we use the real-time 3D human pose tracking by the external smart edge sensors to send semantic feedback informing the robot about people in its vicinity but out of sight of its internal sensors, e.g., due to occlusions or limited FoV (cf. Sec. III-C). This scenario is common in many household or office environments, e.g., at corridor intersections, as well as for warehouses with narrow aisles between high shelves, where people suddenly emerging from behind occlusions can be at risk of collision with autonomously operating robots.

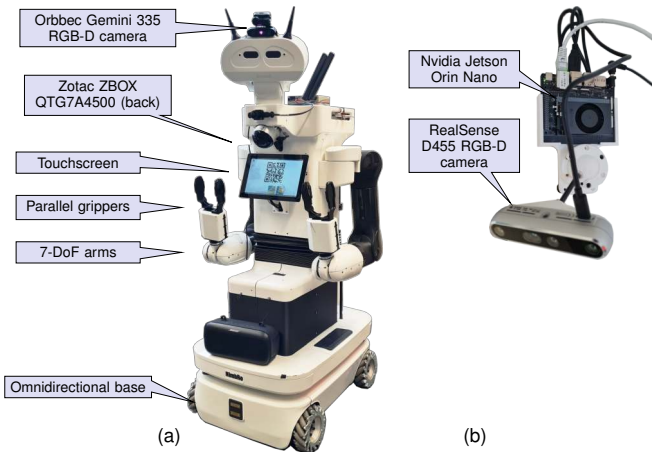


Fig. 6. Robot and sensor setup: (a) PAL Robotics TIAGo++ robot; (b) sample smart edge sensor.

The experiment is illustrated in Fig. 7 for a scenario where a person emerges from behind an occluding wall and crosses the robot’s path. Without feedback (Column (a)), the robot can react to the person only after they emerge from behind the wall and are visible in its own sensors. Robot and person come dangerously close. With semantic feedback from the external sensors about tracked persons (Column (b)), the person with its linear velocity estimate is included in the robot’s dynamic obstacle map, where regions they (prospectively) walk through are marked as occupied (dark gray color) even before they appear from behind the occlusion. The robot adapts its navigation path foresightedly and keeps a safe distance from the person who is crossing its originally planned path.

In Tab. I, we provide a quantitative evaluation of the minimum safety distance towards the human maintained by the robot during five iterations of the experiment with and without anticipation, respectively. The experiment was conducted for two robots, HSR (H) and TIAGo (T), and two subjects (S1 and S2). The HSR experiments were executed only by S1. The experiments show that the minimal safety distance is significantly higher with anticipation compared to the default robot behavior. Through the semantic human pose feedback, the robots can anticipate a person emerging from behind an occlusion significantly earlier and anticipatorily adjust their navigation path to always maintain a sufficient safety distance of at least 82 cm or 50 cm for HSR and TIAGo, respectively. Without anticipation, the robots come dangerously close to the emerging persons with a worst-case distance of only 11 cm resp. 8 cm.

2) *Collaborative Furniture Transport*: In a second set of experiments, we evaluate collaborative furniture carrying with the TIAGo robot. The task consists of collaboratively arranging tables to a user-defined target layout. The experiment is conducted with and without anticipatory control. *With active anticipation*, the smart edge sensor network predicts the human’s intention of which table to lift and from which side of the table to assist the human (cf. Sec. III-C). Furthermore, the compliant control mode is guided by anticipating the goal location of the table placement (cf. Sec. III-D). *Without*

TABLE I
AVERAGE AND WORST-CASE PERSON–ROBOT SAFETY DISTANCE.

	w/o anticipation			w/ anticipation		
	S1–H	S1–T	S2–T	S1–H	S1–T	S2–T
avg.	0.41 m	0.23 m	0.19 m	0.86 m	0.71 m	0.79 m
worst	0.11 m	0.08 m	0.12 m	0.82 m	0.50 m	0.61 m

Results from five runs of the anticipatory navigation experiment with HSR (H) and TIAGo (T) robots for two subjects (S1, S2).

TABLE II
COLLABORATIVE FURNITURE CARRYING PERFORMANCE.

	Anticipation	Trans. Error (m)	Ang. Error ($^{\circ}$)	Duration (s)
✓	S1	0.06 \pm 0.04	0.64 \pm 0.57	88 \pm 11
✓	S2	0.06 \pm 0.04	0.36 \pm 0.47	86 \pm 7
–	S1	0.13 \pm 0.03	6.41 \pm 5.64	106 \pm 9
–	S2	0.29 \pm 0.07	14.50 \pm 10.44	124 \pm 16

Experiment with (top) and without anticipation (bottom). The results of each subject are averaged over five trials.

anticipation, the smart edge sensor network does not provide global context information for anticipation. The robot cannot anticipate the human’s intention but waits for human input on a touchscreen interface. The compliant control mode during table carrying is active and translates to simple robot velocity control, without target location anticipation. Two subjects performed the task with and without anticipation for five repetitions each. Sample visualizations from the semantic scene model and a camera view are shown in Fig. 8.

We evaluate the task performance in terms of translation and angular error, and task completion time. The translation error is calculated as the Euclidean distance between the goal pose and the actual placement pose of the robot. Similarly, we calculate the angular error as the yaw angle difference. The task achievement time is measured from the start of the task until the table is placed at the goal location.

The results are shown in Tab. II. The table carrying task is executed faster with anticipation and the pose error is smaller compared to the task without anticipation. On average, the anticipated collaboration was completed 26 s faster than the user interface-based baseline. Furthermore, the target location is achieved more accurately with anticipation.

C. Collaborative Furniture Rearrangement

Finally, we show the integration of our anticipation approaches to demonstrate their potential in a collaborative task of arranging furniture to a target room layout as depicted in Fig. 9. This demonstration combines anticipatory human-aware navigation and collaborative furniture arrangement, including collaboratively carrying tables and moving chairs. During the 11-minute continuous operation, four tables were carried and two chairs were moved to the previously placed tables. For each table, it was anticipated which one should be carried and from which side, based on the smart edge sensor semantic scene model. In addition, the robot anticipated the desired placement location based on the room layout while compliantly collaborating with the human to carry the table. The expected target pose was automatically updated based on the location in the furniture plan the person was

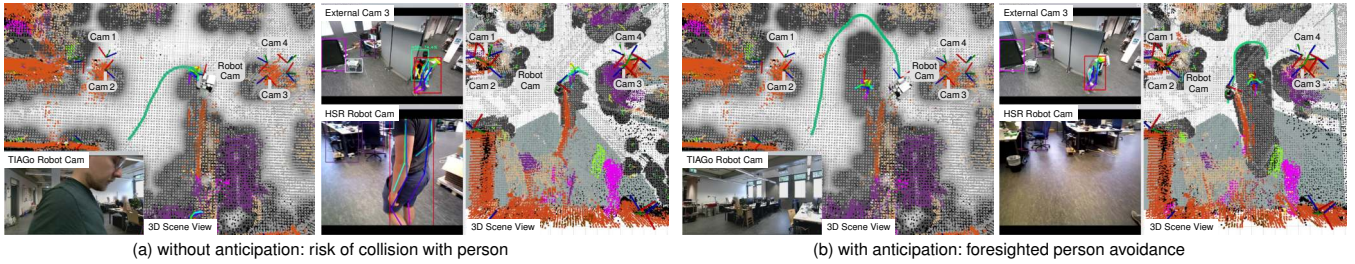


Fig. 7. Human-aware anticipatory navigation: Visualization of a person emerging from behind an occluding wall for TIAGo (left) and HSR (right) robots. Semantic feedback about tracked persons and their velocity (red arrow) allows the robots to anticipate and foresightedly adjust their navigation paths.



Fig. 8. Collaborative table carrying task. The robot anticipates the person’s choice of which table to carry and which side to grasp. The person guides the compliant robot via interaction forces through the table. The robot anticipates the target placement location and takes the lead for precise positioning.

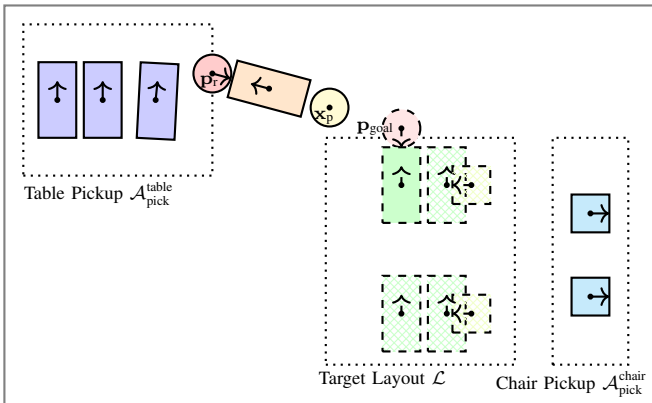


Fig. 9. Collaborative furniture arrangement: Tables and chairs are carried collaboratively with the human from $\mathcal{A}_{pick}^{table}$ resp. $\mathcal{A}_{pick}^{chair}$ to the user-defined target layout \mathcal{L} . Human, robot, and object poses are estimated by the smart edge sensor network and used to anticipate and navigate to the pickup pose (Alg. 1, 2). During collaborative carrying, the robot anticipates the target table pose based on \mathcal{L} (Alg. 3). Chairs can be moved by the robot alone.

approaching. The chairs to align were also identified by the smart edge sensors. When navigating back to the pickup location, the robot received semantic feedback about the predicted movement of persons, also behind occlusions, and anticipatorily adapted its navigation path to avoid collisions and keep a sufficient safety distance. Fig. 10 depicts excerpts of the demonstration, which was conducted with two different subjects. Please also refer to the accompanying video.

V. CONCLUSIONS

In this paper, we presented approaches to incorporate allocentric semantic context information from smart edge sensor network observations to anticipate human behavior on two levels: (1) in the context of human-aware navigation to

improve safety, by projecting future human trajectories into the planning map of a mobile robot, and (2) in the context of collaborative mobile manipulation for improving efficiency, by anticipating intentions to work towards a desired goal.

Both approaches were evaluated in real-world experiments and compared against non-anticipatory baseline approaches utilizing a graphical user interface for human-robot interaction. Our approach demonstrates safer human-aware navigation and improved efficiency for human-robot collaboration with a mobile manipulation robot. We show that the robot anticipates persons emerging from behind occlusions and preemptively adjusts its navigation path to maintain a safe distance by incorporating semantic feedback of human pose observations from external sensors.

An integrated demonstration shows our approach’s potential for collaborative human-robot interaction, achieving the complex task of setting a room layout with tables and chairs.

Directions for future work include implementing anticipatory human-aware navigation also on a higher planning level, instead of using the local obstacle cost map, taking long-term goals and intents of the persons into account.

REFERENCES

- [1] G. Hoffman, “Anticipation in human-robot interaction,” in *It’s All in the Timing*, AAAI Spring Symposium, 2010.
- [2] C.-M. Huang and B. Mutlu, “Anticipatory robot control for efficient human-robot collaboration,” in *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 83–90.
- [3] L. Psarakis, D. Nathanael, and N. Marmaras, “Fostering short-term human anticipatory behavior in human-robot collaboration,” *International Journal of Industrial Ergonomics*, vol. 87, p. 103 241, 2022.
- [4] C. Canuto, P. Moreno, J. Samatelo, R. Vassallo, and J. Santos-Victor, “Action anticipation for collaborative environments: The impact of contextual information and uncertainty-based prediction,” *Neurocomputing*, vol. 444, pp. 301–318, 2021.

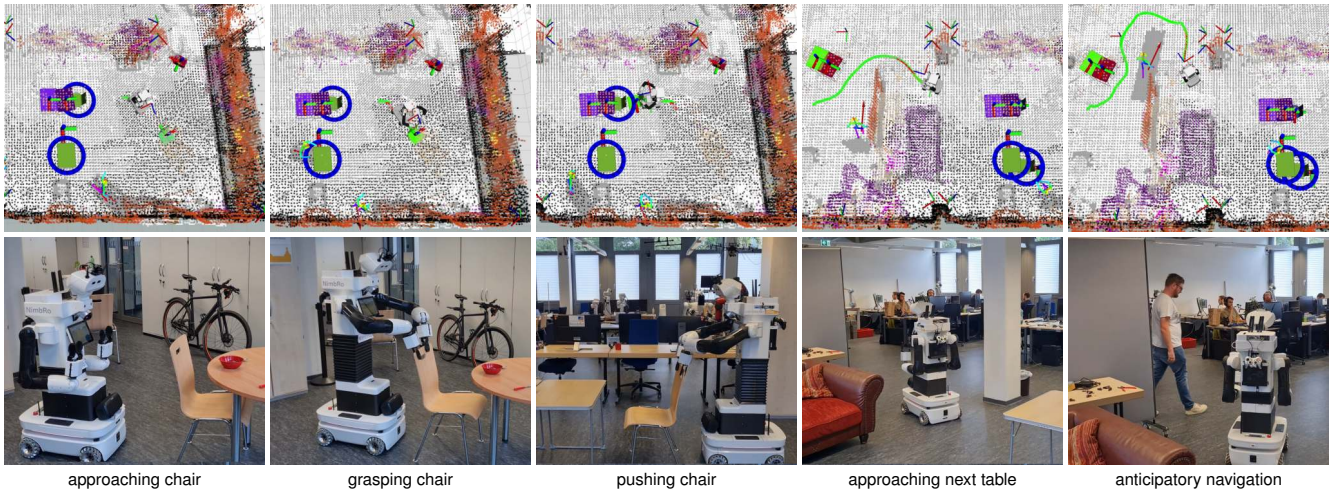


Fig. 10. Visualizations of the collaborative furniture arrangement. Human, robot, and object poses are tracked by the smart edge sensor network. Green markers with blue circles denote the anticipated target pose for the next chair resp. table. The robot autonomously approaches the chair, grasps and pushes it to the goal pose given by the target layout. It then approaches the next table, preemptively adapting its navigation path to the person appearing from behind the occluding wall.

- [5] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: Reading the intentions of humans and robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4132–4139, 2018.
- [6] S. Bultmann, R. Memmesheimer, and S. Behnke, "External camera-based mobile robot pose estimation for collaborative perception with smart edge sensors," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8194–8200.
- [7] J. Tanke, L. Zhang, A. Zhao, C. Tang, Y. Cai, L. Wang, P.-C. Wu, J. Gall, and C. Keskin, "Social diffusion: Long-term multiple human motion anticipation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 9601–9611.
- [8] L. Bärmann, R. Kartmann, F. Peller-Konrad, J. Niehues, A. Waibel, and T. Asfour, "Incremental learning of humanoid robot behavior from natural interaction and large language models," *preprint arXiv:2309.04316*, 2023.
- [9] J. de Heuvel, N. Corral, B. Kreis, J. Conradi, A. Driemel, and M. Bennewitz, "Learning depth vision-based personalized robot navigation from dynamic demonstrations in virtual reality," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2023.
- [10] M. Luber, L. Spinello, J. Silva, and K. O. Arras, "Socially-aware robot navigation: A learning approach," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 902–907.
- [11] L. Bruckschen, K. Bungert, N. Dengler, and M. Bennewitz, "Human-aware robot navigation by long-term movement prediction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 11 032–11 037.
- [12] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, p. 4282, 1995.
- [13] G. Ferrer, A. G. Zulueta, F. H. Cotarelo, and A. Sanfeliu, "Robot social-aware navigation framework to accompany people walking side-by-side," *Autonomous robots*, vol. 41, no. 4, pp. 775–793, 2017.
- [14] P. Arena, L. Fortuna, M. Frasca, and L. Patané, "Learning anticipation via spiking networks: Application to navigation control," *IEEE transactions on neural networks*, vol. 20, no. 2, pp. 202–216, 2009.
- [15] D. Rus, B. Donald, and J. Jennings, "Moving furniture with teams of autonomous robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 1, 1995, pp. 235–242.
- [16] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "IkeaBot: An autonomous multi-robot coordinated furniture assembly system," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 855–862.
- [17] J. Stückler and S. Behnke, "Following human guidance to cooperatively carry a large object," in *11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2011, pp. 218–223.
- [18] J. Stückler, M. Schwarz, and S. Behnke, "Mobile manipulation, tool use, and intuitive interaction for cognitive service robot Cosero," *Frontiers Robotics AI*, vol. 3, p. 58, 2016.
- [19] A. Fallatah, B. Stoddard, M. Burnett, and H. Knight, "Towards user-centric robot furniture arrangement," in *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 1066–1073.
- [20] S. Bultmann and S. Behnke, "Real-time multi-view 3D human pose estimation using semantic feedback to smart edge sensors," in *Robotics: Science and Systems (RSS)*, 2021.
- [21] S. Bultmann and S. Behnke, "3D semantic scene perception using distributed smart edge sensors," in *International Conference on Intelligent Autonomous Systems (IAS)*, 2022, pp. 313–329.
- [22] J. Hau, S. Bultmann, and S. Behnke, "Object-level 3D semantic mapping using a network of smart edge sensors," in *6th IEEE International Conference on Robotic Computing (IRC)*, 2022, pp. 198–206.
- [23] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLOv8*, <https://github.com/ultralytics/ultralytics>, version 8.0.0, 2023.
- [24] NVIDIA Corporation, *NVIDIA Omniverse Replicator*, https://docs.omniverse.nvidia.com/extensions/latest/ext_replicator.html, 2024.
- [25] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.
- [26] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International Journal of Computer Vision (IJCV)*, vol. 81, pp. 155–166, 2009.
- [27] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *International Journal of Computer Vision (IJCV)*, vol. 109, no. 3, pp. 252–286, 2014.
- [28] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.
- [29] X. Zhao, Y. Chen, S. Xu, X. Li, X. Wang, Y. Li, and H. Huang, "An open and comprehensive pipeline for unified object grounding and detection," *preprint arXiv:2401.02361*, 2024.
- [30] NVIDIA-AI-IOT, *NanoSAM*, <https://github.com/NVIDIA-AI-IOT/nanosam>, 2024.
- [31] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight SAM for mobile applications," *preprint arXiv:2306.14289*, 2023.
- [32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.
- [33] J. Pages, L. Marchionni, and F. Ferro, "TIAGo: The modular robot that adapts to different research needs," in *International Workshop on Robot Modularity, IROS*, vol. 290, 2016.