IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 July 2025, accepted 29 August 2025, date of publication 9 September 2025, date of current version 16 September 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3607191*

SURVEY

# Leveraging Pedestrian Detection and Tracking in Robotics Navigation: A Survey With Practical Illustrations

**N. PICELLO[1], F. HERRERO[1], S. HERNÁNDEZ[2], A. LÓPEZ[2], AND A. SANTAMARIA-NAVARRO[1]**

[1]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Universitat Politècnica de Catalunya (UPC), 08028 Barcelona, Spain
[2]Consejo Superior de Investigaciones Científicas (CSIC), Institut de Robòtica i Informàtica Industrial (CSIC-UPC), 08028 Barcelona, Spain

Corresponding author: N. Picello (npicello@iri.upc.edu)

This work has been partially supported by Horizon Europe under the grant agreement No. 101168042 (TRIFFID: auTonomous Robotic aId For increasing FIrst responDers efficiency); by the Spanish Ministry of Science and Innovation under the project "LENA: Lifelong navigation learning using human-robot interaction" (PID2022-142039NA-I00, funded by the "Ministerio de Ciencia, Innovación y Universidades" through the "Agencia Estatal de Investigación", MCIN/AEI/10.13039/501100011033 and by the European Regional Development Fund, "ERDF A way of making Europe"); by the project "BotNet: Nou model de repartiment de paquets en superilles urbanes mitjançant una xarxa de vehicles elèctrics autònoms" (23S06128-00), funded by the "Ajuntament de Barcelona" and "Fundació la Caixa"; and by the Consolidated Research Group "RAIG: Mobile Robotics and Artificial Intelligence Group" (2021 SGR 00510) of the "Departament de Recerca i Universitats de la Generalitat de Catalunya".

**ABSTRACT** Pedestrian Detection and Tracking (PDT) plays a pivotal role in enabling autonomous robots to navigate safely and efficiently in dynamic, human-populated environments. This paper presents a comprehensive survey of PDT methods, structured according to the sensing modalities employed: RGB cameras, LiDAR, thermal imaging, RGB-D sensors, and multi-modal fusion systems. For each category, we analyze representative techniques, synthesize their strengths and limitations, and discuss recent advancements including deep learning approaches and cross-modal fusion strategies. We highlight persistent challenges such as handling occlusions, achieving real-time performance, and ensuring robustness across diverse environments. In addition to this structured review, we provide two practical examples using the Ona autonomous robot platform (see Figure 1) to illustrate how PDT techniques can enhance robotic capabilities in real-world scenarios. These examples focus on improving SLAM consistency and enabling proxemic-aware navigation strategies. Through this survey, we aim to clarify the current state of the art, identify emerging trends, and suggest future research directions for robust and socially-aware robotic navigation.

**INDEX TERMS** Human-aware navigation, multi-sensor fusion, pedestrian detection, robot navigation, people tracking.

## I. INTRODUCTION

Autonomous robots are playing a growing role in society, operating in environments frequently shared with humans. As robots begin to navigate shared spaces-from public venues to private residences-it is essential for them to possess robust capabilities for localizing people in their vicinity. Accurate human localization is critical, not only for enabling effective and natural human-robot interactions but also for ensuring safe coexistence. Robots equipped with precise human-awareness can better interpret social cues, avoid collisions, and operate seamlessly alongside people, significantly enhancing their utility and acceptance across diverse application domains [1], [2]. Robots should be able to navigate with and around people in a socially acceptable and efficient way, making traditional navigation algorithms, which focus solely on finding the shortest path and treat humans as simple dynamic obstacles, obsolete [3].

In this context, the task of Pedestrian Detection and Tracking (PDT) emerges as a key enabling technology for autonomous robots and vehicles. PDT refers to the detection, localization, tracking, and sometimes prediction of human movements in the robot's surroundings. These capabilities are crucial for enabling robots to maintain continuous awareness

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Abdur Razzaque.

of their environment, to anticipate human actions, and to navigate dynamically and safely through shared spaces. Despite significant advances in recent years, PDT remains a challenging problem due to a wide range of factors: sensor limitations, dynamic and cluttered environments, frequent occlusions, and the inherent unpredictability of human behavior. Researchers have proposed a broad spectrum of approaches for PDT, leveraging different sensing modalities such as RGB cameras, LiDAR, RGB-D sensors, thermal cameras, and multi-modal fusion. This paper provides a comprehensive and structured survey of these methods, analyzing their strengths, weaknesses, and applicability to robotic navigation. In particular, we organize this review according to sensing modality and critically assess each category in terms of detection performance, robustness to challenging conditions, and computational demands.

Real-world applications highlight the necessity of robust PDT methodologies, especially in diverse environments beyond conventional roads, including crowded public spaces, airports, hospitals, commercial zones, and urban pedestrian areas. In these contexts, frequent human-robot interactions demand high standards for operational efficiency and safety. Beyond surveying the state of the art, we also illustrate how PDT can concretely benefit robot navigation through two practical examples using the Ona autonomous platform, shown in Figure 1. These examples are not presented as core contributions but serve as case studies demonstrating how established PDT techniques can improve navigation performance in realistic settings. The first example shows how PDT facilitates cleaner and more reliable SLAM mapping by effectively mitigating the negative influence of dynamic elements on map generation, allowing us to successfully identify and remove 3D sensed points belonging to people from a static 3D mapping point cloud of the robot's surroundings, meant for robot localization. The second example demonstrates the integration of pedestrian tracking into a proxemic-aware navigation framework, enabling robots to respect personal space and anticipate human motion for safer, socially-aware planning. These examples underscore the significant benefits of PDT for robots operating in environments characterized by continuous pedestrian activity, which traditionally complicates accurate mapping and efficient path planning.

In the next section, we provide an exhaustive literature review on pedestrian detection and tracking, presenting a classification of the various PDT approaches one can explore based on different sensor modalities (i.e., observation models). Section III outlines the advantages of accurately localizing pedestrians and Section IV describes how such information can significantly enhance robot navigation performance through real-world experiments. Finally, conclusions are drawn in Section V.

## II. LITERATURE REVIEW

PDT has evolved significantly over the last two decades, driven by advances in sensing hardware and learning-based algorithms. This section organizes the current state-of-the-art according to sensing modality: vision-based (RGB), LiDAR-based, RGB-D, thermal, and multi-modal systems. Within each category, we compare detection and tracking techniques, discuss their strengths and limitations, and highlight recent trends and gaps. Each observation model, summarized in Table 1, presents trade-offs in terms of accuracy, computational cost and environmental adaptability. As a result, sensor fusion remains a dominant trend in pedestrian detection and tracking, maximizing robustness and reliability across different operating conditions.


**FIGURE 1.** ONA, the robot used in our experiments.

### A. VISION-BASED METHODS (RGB)

Early vision-based methods relied heavily on handcrafted features and classical classifiers. Histograms of Oriented Gradients (HOG) combined with Support Vector Machines (SVM) were among the earliest successful methods for pedestrian detection [4], [5]. For instance, [7] used HOG and Local Binary Patterns (LBP) for detection, with Kalman filtering for tracking. Similarly, [4] proposed a cascade of boosted classifiers for faster inference. These methods, although foundational, suffer from sensitivity to lighting and occlusion. Reference [6], improved the detection robustness by combining sparse-stereo ROI extraction, shape-based detection, and texture classification using neural networks. The tracking task was performed leveraging the Hungarian algorithm in [6]. In later works, Aggregate Channel Features (ACF) [8], [9], [10] provided a strong baseline for pedestrian detection before the deep learning era. The initial use of ACF later evolved into hybrid frameworks integrating deep learning to enhance detection accuracy [11], [12].

The rise of CNNs dramatically improved detection performance. Detectors such as YOLO [13] and Fast R-CNN [14] offered real-time, high-accuracy detection even in cluttered environments. Coupled with tracking-by-detection methods like Deep SORT, recent approaches using YOLOv5 and YOLOv7 have enhanced data association accuracy [15], [16]. However, RGB-based systems remain limited by their lack of depth perception and susceptibility to occlusion.

**TABLE 1. Comparison of observation models used in pedestrian detection and tracking systems.**

| Observation Model | Advantages | Disadvantages | Works |
|---|---|---|---|
| Vision-based | - Rich visual data for object recognition<br>- Compatible with deep learning models<br>- Lightweight, low power, low cost<br>- High resolution and fast with optimized models | - No spatial/depth information<br>- Limited field of view<br>- Sensitive to lighting, weather, and occlusions<br>- Prone to motion blur | [4]–[20] |
| LiDAR-based | - Accurate spatial/depth information<br>- Robust to lighting and weather<br>- Wide field of view<br>- Effective in diverse environments | - No visual context<br>- High computational cost<br>- Expensive hardware<br>- Typically lower frame rates | [21]–[26] |
| RGB-D-based | - Combines visual and depth information<br>- High-resolution input<br>- Good for indoor and short-range tasks | - Sensitive to sun light and reflective surfaces<br>- Requires accurate calibration<br>- Limited range and field of view | [20], [23], [27]–[32] |
| Thermal-based | - Performs well in low-light or night conditions<br>- Highlights warm-bodied objects (e.g., humans)<br>- Lightweight and passive sensor | - Sensitive to environmental heat variations<br>- Limited object class discrimination<br>- Narrow field of view | [27], [33]–[37] |
| Hybrid-based | - Fuses complementary sensor data<br>- Increases robustness and accuracy<br>- Better generalization across scenarios | - Requires precise calibration and synchronization<br>- High complexity and cost<br>- Potential latency in fusion | [38]–[48] |

To address such challenges, [17] proposes a saliency-aware tracker that jointly reasons about human attention and motion cues to improve temporal consistency. Similarly, [18] introduces a language-grounded visual tracker that incorporates natural language specifications, enabling context-aware pedestrian tracking in complex scenes. These methods suggest a promising direction where multi-modal semantic understanding enhances classical vision-based PDT.

In terms of tracking strategies, Kalman Filtering remains a prevalent technique for pedestrian state estimation [19], often used alongside CNN-based detectors [20] and multi-pedestrian tracking is basically achieved by identifying the same pedestrians across frames based on motion and appearance.

### B. LIDAR-BASED METHODS

LiDAR-based pedestrian detection methods leverage 3D geometric features, offering accurate spatial localization and robustness to illumination. For instance, [21] filters and projects 3D point-cloud data onto a 2D occupancy grid, clusters it into blobs, and classifies pedestrian candidates using an RBF-SVM, followed by Kalman filtering for pose prediction. Approaches such as [22] use region-of-interest (ROI) mechanisms to down-sample point clouds, followed by classification using handcrafted features (shape, normals and shade). Others like [23] employ convolutional autoencoders and connected-component algorithms to segment and track pedestrians in 3D LiDAR data.

Unlike image-based systems, LiDAR methods handle cluttered environments and dynamic occlusions more effectively. However, they are computationally expensive and often require specialized point cloud processing pipelines. Techniques such as Kernel Density Estimation (KDE) [24] and Doppler LiDAR [25] have further improved tracking by enhancing pedestrian segmentation and estimating velocity profiles. Reference [26] proposes a quality-aware 3D tracker with shape completion, enabling better pedestrian association even when partial occlusions occur in sparse LiDAR scans. This approach bridges a known gap in traditional 3D tracking pipelines, particularly for long-range or edge-of-sensor scenarios.

Further, using 3D LiDAR sensors also implies lower update rates (10-20 Hz). Alternatively, 2D laser scanners enable faster processing due to the nature of the 2D point-cloud they capture (consisting of way fewer points with respect to its 3D counterpart), but often require integration with other sensors for reliable pedestrian detection. For example, [41] integrates LiDAR and RGB data with optical flow estimation for collision avoidance.

### C. RGB-D METHODS

RGB-D sensors offer a compelling trade-off between appearance and depth, commonly used in indoor scenarios. Affordable sensors like the Microsoft Kinect[1] or Real Sense[2] depth cameras have popularized their usage in robotics.

In PDT, RGB-D cameras enhance scene segmentation and spatial relation assessments. For example, [27] proposes a three-stage cascade method that transforms RGB-D data into a Point Ensemble Image (PEI) for unsupervised detection and classification, followed by trajectory generation. Similarly, [28] use tracking-by-detection on 3D point clouds constructed from the depth data, while [29] classify 3D points into fixed structures, ground, and objects, followed by depth-based upper-body detection and tracking using an Extended Kalman Filter (EKF). Additionally, [30] employs u-depth and v-depth maps for obstacle detection and tracking, while [20], [23] integrates YOLO for RGB-based segmentation and depth data for positional tracking. Despite the advantages of incorporating depth to RGB images, RGB-D sensors have a narrow field of view and use infrared light, which is

[1]https://azure.microsoft.com/en-us/products/kinect-dk
[2]https://www.intelrealsense.com/

affected by the existence of sunlight and thus limits pedestrian localization in most outdoor environments.

Works like [20], [40] integrate leg detection from 2D LiDAR with human pose estimation from RGB-D cameras to improve reliability. Even though recent methods have improved depth-based detection pipelines by fusing 2D image features with 3D spatial cues, RGB-D systems remain limited by narrow field-of-view and sensitivity to sunlight or reflective surfaces. Reference [32] addresses resolution limitations in RGB-D systems by proposing a multi-scale structure-enhanced super-resolution framework. It improves the clarity of low-resolution pedestrian data captured by budget sensors, boosting detection rates in dense indoor settings.

### D. THERMAL-BASED METHODS

Thermal cameras capture infrared radiation, making them effective for low-light conditions where traditional cameras fail. These models typically use shape and appearance features for detection and tracking. Classical works such as [27] combine HOG and DCT (Discrete Cosine Transform) descriptors for thermal ROI classification. Similarly, [33] uses ROI extraction and classification with SVM trained on HOG and DCT features, followed by optical flow-based tracking. Reference [34] proposes a joint shape-and-appearance-based method combined with shot segmentation for tracking, formulated as a weighted bipartite graph matching problem. Recent approaches leverage deep learning; for instance, [35] uses two RetinaNet [36] models to process thermal images and saliency maps, combining features for classification and regression. While tracking is not explicitly addressed in this work, standard algorithms can be integrated post-detection.

Thermal cameras are particularly advantageous in dark environments but may struggle with temperature variations and occlusions. Further, existing commercial thermal cameras are of very low resolution. Hence, thermal-only systems suffer from low resolution and false positives due to environmental heat sources. To overcome this, [37] introduces a cross-modality proposal-guided feature mining method that jointly learns features from registered and unregistered RGB-thermal pairs. The fusion enhances detection reliability in variable lighting.

### E. MULTI-MODAL AND SENSOR FUSION APPROACHES

Combining multiple sensing modalities increases robustness and there is an increasing trend in integrating hybrid approaches that combine multiple sensor modalities. For instance, [38] fuses leg and face detections obtained from LiDAR-based and RGB-based observations, respectively, using a sequential Unscented Kalman Filter (UKF). Similarly, [39] integrates laser scans with stereo vision, employing polyline-based feature extraction and pattern recognition for accurate position estimation and pedestrian identification. Tracking and position prediction are achieved through statistical validation gates and confidence regions.

Other hybrid methods combine 2D LiDAR with RGB-D data, like [40] in which they process 2D LiDAR scans to detect pedestrian leg positions, followed by RGB-D data for human skeleton pose detection. Here, tracking is performed using a Kalman filter with global nearest-neighbor data association. Additionally, [41] fuses 2D LiDAR and RGB imagery, employing optical flow estimation and object detection to compute obstacle and pedestrian positions and velocities.

Thermal and visible-light camera fusion is another effective hybrid strategy. Works such as [42], [43], and [44] exploit thermal and visible spectra for robust pedestrian localization under varying environmental conditions. Deep learning-based hybrid approaches are also gaining traction. For example, [45] converts LiDAR data into depth images, which are processed alongside RGB images through deep neural networks for pedestrian detection. Tracking combines Kalman filter predictions with optical flow techniques. Finally, [46] demonstrates a unique fusion of LiDAR and millimeter-wave radar, where radar detects moving obstacles via Doppler shifts, and LiDAR localizes and segments them, improving dynamic obstacle handling in SLAM scenarios.

Fusion methods face challenges in sensor synchronization, calibration, and real-time processing. However, they yield more robust performance in complex scenes. Notably, [47] and [48] demonstrate that fusing vision and language in a time-evolving latent state improves contextual understanding and long-term tracking, even under partial observability. Overall, a key trend across modalities is the shift towards hybrid and cross-modal architectures, where learning-based fusion outperforms handcrafted pipelines. Nevertheless, real-time constraints, occlusion handling, and transferability across environments remain open challenges.

## III. ADVANTAGES OF DETECTING AND TRACKING PEOPLE

By integrating PDT into navigation systems, robots can better anticipate collisions, adjust their paths dynamically, and interact with pedestrians in a socially acceptable manner.

A key advantage of PDT is its ability to enhance situational awareness. Accurate pedestrian localization allows robots to react in real-time to human motion patterns, reducing the risk of collisions. Studies such as [12] highlight the importance of real-time pedestrian tracking for human-aware navigation. Tracking modules further improve reliability by smoothing detection noise and handling short-term occlusions using probabilistic models such as Kalman Filters and IMM filtering [20], [23], [33]. Related to this, a key aspect of path planning in crowded environments is the ''freezing robot problem'' that occurs when the planner perceives no safe path and the robot is thus forced to stop dead. In a PDT pipeline, the tracking component helps mitigate this issue by enabling the robot to anticipate pedestrian motion. Furthermore, assuming a certain level of human cooperation can enhance navigation even more. As explained in [49], in crowded environments where people naturally adjust their

trajectories to avoid each other, robots can leverage tracked pedestrian paths and improve their navigation success by assuming that humans will also make room for them in shared spaces.

Despite these benefits, PDT remains challenged by occlusions, crowded scenarios, and the computational demands of maintaining real-time awareness in unstructured environments. These limitations, noted across many works in the literature [23], [25], [26], [37], directly impact the reliability of downstream tasks such as navigation and interaction. Addressing these issues remains a key priority for future PDT research.

Beyond detection, PDT plays a crucial role in socially-aware motion planning. In crowded environments, global and local planning must take into account human presence. Treating people as mere moving obstacles does not suffice and a good planner should also consider the personal space and visibility (line-of-sight) of each person, in order to keep a safe distance from people, and comply with general social rules, such as approaching a person from behind [50]. In addition, Robots that actively detect and track humans can infer pedestrian intentions, such as whether a person is walking towards or away from them.

By leveraging pose estimation techniques and velocity tracking [40], robots can adjust their movement in ways that align with human expectations, improving social acceptance in shared spaces. Other social norms such as keeping to one side or not barging through a conversation should be taken into account too, as done in [51], where the authors exploited reinforcement learning to teach the robot which social rules it should follow. In addition to planning around predicted motion, PDT also contributes to the legibility of robot behavior, i.e., making its intentions clear to nearby humans. Furthermore, PDT enables richer reasoning over social contexts when fused with additional modalities, such as language grounding [18] or attention modeling [17]. These extensions offer pathways toward more socially-intelligent navigation strategies, although practical deployment remains limited by current computational constraints.

In addition to immediate (short-term) avoidance, PDT also enables proactive path planning based on predicted pedestrian trajectories. Works such as [41] demonstrate how integrating pedestrian motion prediction into velocity-based planners allows robots to select safe, collision-free paths in dynamic environments. In recent years, the advent of Long Short-Term Memory (LSTM) ( [52]) and Generative Adversarial networks (GAN) ( [53]) have further propelled this PTD field. By using a history of a pedestrian's motion, these models predict where a person will be in the immediate future. An example of this technique is [47], where an LSTM-based predictor is integrated so that the robot can "imagine" where up to 5 bystanders will move, therefore planning safe paths accordingly. Instead, [54] combines Recurrent Neural Networks (RNN) and Mixture Density Networks (MDN) into a Probabilistic Crowd GAN, where the generator's MDN solves for probabilistic multi-modal predictions from which likely modal paths are found (which are compared with the ground truth by the discriminator in the adversarial training process).

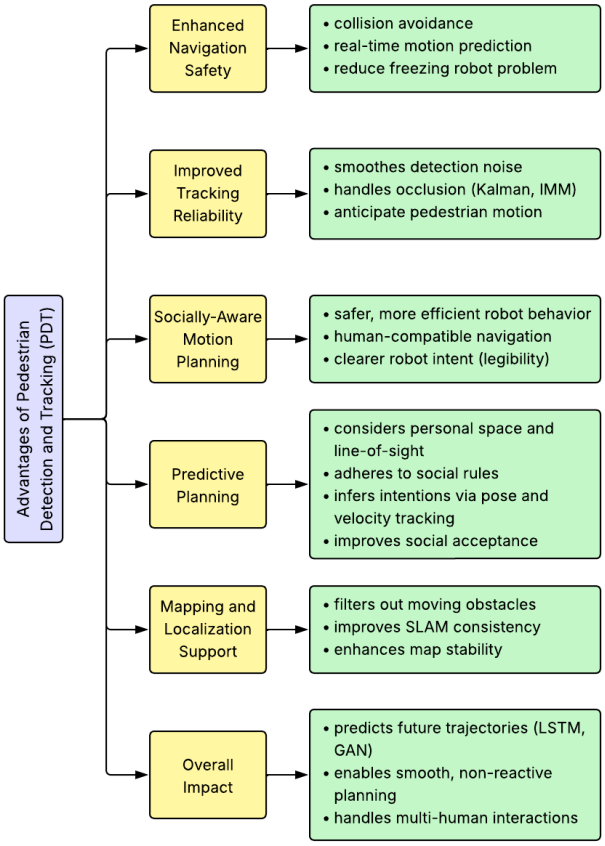



**FIGURE 2. Advantages of leveraging pedestrian detection and tracking in the navigation pipeline.**

PDT also assists mapping and localization tasks by filtering out moving obstacles from the environment model. Traditional SLAM systems struggle with dynamic objects, leading to mapping inconsistencies [20], [23]. By segmenting and removing pedestrians from the mapping pipeline (i.e., disregarding their point cloud data) robots can generate more stable maps, therefore improving localization accuracy. Emerging works such as [32] also demonstrate how enhancing the quality of sensing data, e.g., through super-resolution techniques, can indirectly benefit PDT-driven SLAM pipelines in low-resolution or challenging conditions.

In summary, the integration of pedestrian tracking into robotic navigation systems enhances safety, efficiency and human compatibility. The diagram in Figure 2 summarizes the key concepts of leveraging PDT on the navigation stack of a robot. Whether for dynamic obstacle avoidance, socially-aware behavior or improved mapping, PDT remains a fundamental component in enabling robots to operate seamlessly in human-populated environments.

## IV. PRACTICAL EXAMPLES ON HOW TO LEVERAGE PDT IN THE NAVIGATION OF A ROBOT EQUIPPED WITH A MULTI-MODAL SENSORY SUITE

In this section, we provide two illustrative examples demonstrating how Pedestrian Detection and Tracking (PDT) techniques can be integrated into the navigation pipeline of a robot equipped with a multi-modal sensory suite. These examples serve to contextualize the methodologies reviewed earlier and show their practical relevance in real-world scenarios. They are presented not as novel contributions but as application vignettes that align with established findings in the literature on the benefits of integrating PDT into navigation.

### A. ROBOT SETUP

The robot used in our experiment is ONA [55], shown in Figure 1. Ona is an all-electric, last-mile autonomous delivery vehicle with an autonomy of more than 5 hours of continuous operation. It weighs around 200 kilos and is roughly $1.8 \times 1.1 \times 1$ meters considering the outer shell. Ona is also equipped with six wheels: four front, traction-only, and two rear, traction and steering, wheels. Further, it exploits the following sensors:

- Wheel encoders.
- Inertial Measurement Unit (IMU), providing an estimate of its (3)-axis) angular velocity and (3)-axis) linear acceleration.
- Global Navigation Satellite System (GNSS), to provide global positioning.
- 3D LiDARs: two 3D lasers of 16 beams are installed in the opposite corners of the robot (front-right and back-left corners).
- RGB Cameras: two RGB cameras are mounted at the front and rear sides of the vehicle.
- Bumpers, as last resort hardware safety feature.

In the following experiments we have a ROS-based[3] navigation pipeline. We couple Fast-LIMO[4] with Cartographer ( [56]) to perform Localization and Mapping, respectively. Pre-processing steps must be carried out to refine the raw data captured by the sensors. Specifically, we first merge the point clouds captured by the two LiDARs into a single, unified, cloud. Next, following common practices in autonomous driving applications, we perform a one-shot traversability analysis to remove ground points from the point cloud, so to exclude them from further analysis. To achieve this, we employ our probabilistic graph-based ground segmentation algorithm, as outlined in [57]. Instead of using dedicated laser sensors to produce a 360-degree scan around the robot, we take a simpler approach by picking the nearest point to the robot in each angular bin from the merged point cloud. This effectively generates a planar scan of the surrounding environment, enabling us to use laser-based detection methods on the resulting data. Additionally, the RGB image data is processed by the YOLO-detection architecture to return the bounding boxes, and corresponding centroids, of the pedestrians in the surrounding. After these pre-processing steps and once the sensing data has been de-noised, we are ready for the PDT.

[3]https://www.ros.org

[4]https://github.com/fetty31/fast_LIMO

### B. SPENCER PEOPLE TRACKING

The Spencer People Tracking (SPT) framework ( [58], [59]) was selected for these examples due to its established robustness in multi-sensor environments and its alignment with PDT approaches discussed in the literature. Our adaptations are minor and practical, focused on integrating available sensory data streams rather than proposing methodological innovations. SPT robustly integrates data from multiple sensor modalities to identify and continuously track individuals within dynamic environments. Through a chain-merging mechanism, the framework effectively combines diverse sensor inputs to produce reliable pedestrian trajectories. Our approach to detecting and tracking people builds upon this system, which was originally designed to integrate multiple RGB-D and 2D laser detectors into a unified framework. In our pipeline, however, we adapt this framework to a different set of sensors: 3D point clouds, 2D laser scans, and RGB images. Each sensor stream is processed separately to extract relevant detections, which are then passed to the SPT system.

- 3D Point-cloud detections: After merging the two 3D point clouds and removing the ground plane, we cluster the remaining points to identify potential person-associated clusters. Given the vertical sparsity of such sensors, our clustering algorithm tends to produce outliers, making necessary additional refinement steps by leveraging data from other sensors, hence improving the detection quality.
- 2D scan detections: As an alternative detector using the unified LiDAR point cloud, we extract a 2D laser scan and processes it using a simple blob detector that identifies regions likely to correspond to people [60]. This approach has proven alternative to the 3D point-cloud clustering although it still produces a number of outliers that we need to disregard.
- RGB-based YOLO detections: One of the most innovative elements of our implementation is the integration of YOLO [13] within the Spencer pipeline. The robot's two RGB cameras provide front and rear views of the environment, which are then analyzed by YOLO (i.e., YOLOv11) to detect key objects. We treat YOLO as a ''high confidence'' detector, using its accurate results to help filter out outliers of the previous point cloud processors. The main limitation, however, is the absence of distance (depth) information, which means YOLO detections are valuable for identifying objects but less useful for precise localization. Also, in order to detect people with YOLO, pedestrians must stay within the field of view of the two cameras, which can eventually be limiting when you want to detect and track all the people

in the surroundings. Thus, we will rely on LiDAR-based observations when pedestrians appear out the cameras' field-of-views.

After obtaining all detections from all the modalities we convert them into a standard format and fuse them to initialize or update existing tracks. Here, two fusion mechanisms are employed:

- Euclidean Fuser: Merges multiple detections by calculating the Euclidean distance between composite detections in the X and Y dimensions. It then computes a joint position by averaging the X, Y, Z coordinates (and corresponding covariances) of the two detections.
- Polar Fuser: This method merges multiple detections using polar coordinates. The fused pose is determined by the polar distance between composite detections. Since the Polar Fuser is applied only to YOLO detections, which lack localization (camera depth) data, the final pose relies on other sensors fused with the YOLO outputs.

The tracking component utilizes the Extended Nearest-Neighbor (ENN) Tracker outlined in [59]. This method leverages a nearest-neighbor (NN) data association strategy to match new sensor detections to existing tracks efficiently. In practice, the greedy NN algorithm identifies the closest detection for each existing track, updates the track state, and iterates continuously, ensuring robust performance, especially when multiple sensor modalities are employed.

The tracker incorporates a track initiation logic to mitigate false-positive track formations and rapidly establish reliable tracks when new individuals enter the scene. To further bolster tracking performance, particularly for predicting human movements, an Interacting Multiple Models (IMM) filter is integrated. The IMM filter combines three distinct motion models:

- Constant Velocity (CV) model: Effective for predicting steady pedestrian motion at near-constant speeds.
- Coordinated Turn (CT) model: Handles smooth, curved pedestrian trajectories.
- Brownian Motion (Wiener Process) model: Captures irregular pedestrian movements, including sudden stops, abrupt direction changes, or stationary states.

Integrating these models significantly enhances the robustness and accuracy of the tracking system across diverse pedestrian behaviors and dynamic environmental conditions.

### C. EXAMPLE 1: REMOVING PEDESTRIANS FROM POINT CLOUD

This first example illustrates how existing PDT techniques can improve Simultaneous Localization and Mapping (SLAM) robustness by removing dynamic elements such as pedestrians from LiDAR point clouds. The rationale aligns with prior works such as [46] or [26], and demonstrates the value of filtering dynamic objects to achieve cleaner maps and better localization.

Current autonomous robots rely on accurate SLAM to navigate and interact effectively with their environment. However, despite continuous advancements in SLAM methodologies, the accuracy and reliability of these algorithms remain highly dependent on environmental conditions. First of all, most SLAM algorithms assume a static scenario, so moving objects violate this assumption and can cause mapping errors or localization drift [46]. Unfiltered moving pedestrians often leave "ghost" artifacts in accumulated point-clouds or contribute spurious features that may confuse scan matching. This results in degraded SLAM performance, with maps becoming inconsistent and worse robot-pose estimation.

The problem of removing dynamic obstacles from point-clouds can be tackled in different manners, depending on the processing strategy used within the SLAM framework, such as online real-time filtering, post-processing after the mapping, or long-term SLAM. In our set-up we investigated the early removal of dynamic points from the point-cloud, so to pass the already filtered sensor data to the SLAM algorithm. Reference [46] demonstrates how this procedure yields cleaner maps and more reliable loop closures, and how they managed to reduce the localization error by 30% or more in highly dynamic environments.

In this example, we propose to leverage the detection and tracking of the SPT together with a *point remover* node that filters out their associated points, so as to remove the negative effect in SLAM of dynamic points belonging to pedestrians. To do so, we investigated three different methodologies: Cluster-based; KD-tree-based; and cylinder-based removal. Each method achieved comparable results, but each has its own strengths and weaknesses:

- Cluster-based: Using the clusters identified in previous nodes, we label points in the cloud based on their associated cluster ID. By checking the cluster ID for each detection and determining the nearest cluster centroid, we can identify and remove all points belonging to that cluster. This methodology is the least robust to outliers, since the presence of a centroid may remove big parts of the point-cloud that do not belong to people. Hence the necessity to have a precise tracker for pedestrians.
- Kdtree-based: Starting from the centroid of a person's detected position, we use a KD-tree search to find points within a certain radius and remove them from the point cloud. However, this approach has the drawback of removing points in a spherical region, which can sometimes eliminate nearby points that should remain.
- Cylinder-based: This is the simplest and arguably the most effective approach. In this method, people are treated as cylindrical volumes, and any points falling inside the cylinder are filtered out. Because the ground plane is already removed, this technique only considers the X and Y coordinates, making the method less sensitive to variations in height.

To show the validity of our example, we compare the occupancy grid maps obtained from projecting to a 2D plane

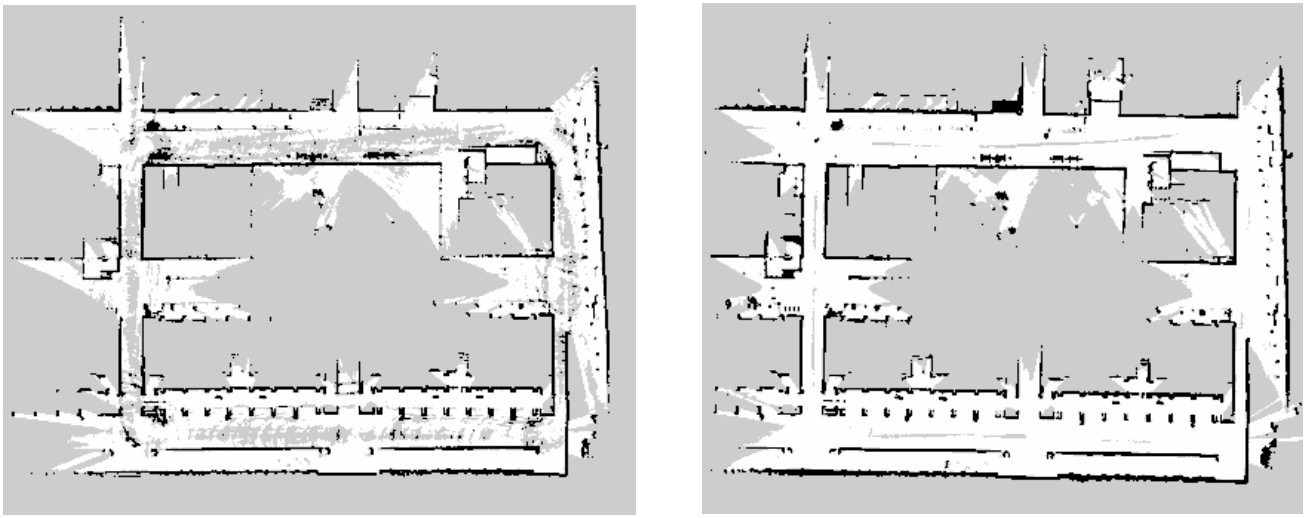


**FIGURE 3.** Comparison of Occupancy grid maps generated. Left: Occupancy grid map generated from the original point cloud (i.e., unfiltered LiDAR scans). Right: Occupancy grid map generated from the processed point-cloud, without considering the points we clustered as belonging to people.

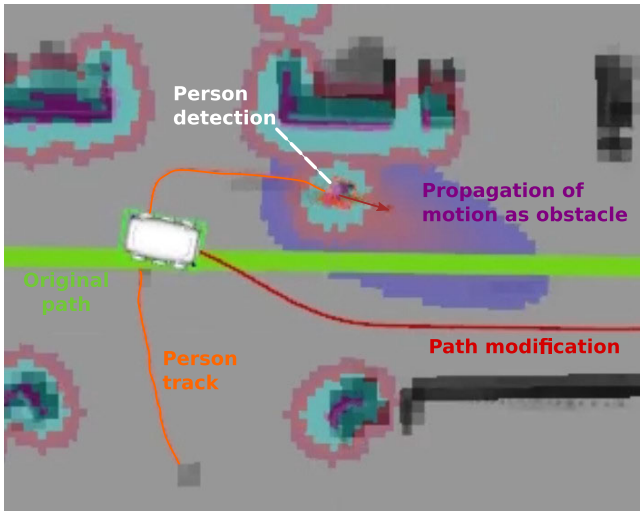



**FIGURE 4.** Occupancy grid map (gray-scale background) with overlapped local cost map (red-blue-scale obstacles) and the original global path (green); the detected person (orange path); and the modified path (red) using the proxemic layer to predict the trajectory of the person as an obstacle in the local path planner.

the resulting map of the SLAM system with and without removing the points belonging to pedestrians. In particular, we can observe map consistency, clarity in narrow passages and reduction in artifact noise as key indicators of improved SLAM performance when leveraging PDT. We tested the implementation in the Barcelona Robot Lab, an outdoor scenario within the North Campus of the Universitat Politècnica de Catalunya[5] [61], which encompasses an outdoor pedestrian area of 10.000 sq m. and is provided with 21 fixed cameras, full coverage of wifi and partial GNSS coverage. The area has moderate vegetation and intense cast shadows, making computer vision algorithms challenging. Further, it is freely occupied by the campus students. In terms of occupation, the density of population in these examples is high enough to interfere with the navigation of the robot although without producing dead ends or blocking situations.

Figure 3 shows a comparison of the 2D occupancy grid maps (OGM) obtained inside the Campus Nord of UPC, generated using the unfiltered point cloud captured by the LiDARs (left); and using the processed version that excludes the points using PDT (right). Considering that these OGMs are then commonly exploited for path planning, the enhancement in the processed OGM is quite clear (right in Figure 3). For instance, we can visually notice a significant reduction in the number of grey-colored points in the occupancy grid map, which correspond to unknown areas. The processed maps exhibit clearer and more defined pathways, especially in narrow streets, with fewer outliers (in dark-grey/black). These outliers are typically caused in the original point cloud by individuals remaining stationary longer than Cartographer's dynamic obstacle filtering can effectively manage. This noise reduction contributes to a more accurate and reliable environmental representation. Consequently, it decreases the complexity of optimizing the motion path in the robot's local planner. Note that limitations in our example include the reliance on prior pedestrian detections from the SPT framework, which may not generalize to all sensor modalities or more challenging environments with high occlusions. The choice of removal strategy (e.g., cluster vs. cylinder-based) also reflects practical trade-offs rather than optimized solutions.

### D. EXAMPLE 2: PEOPLE TRAJECTORY PREDICTION

The second example demonstrates how integrating tracked pedestrian trajectories into a proxemic-aware local planning layer can facilitate socially-aware navigation. This aligns with concepts reviewed in Section III, including adaptive proxemics [62], [63] and trajectory-based planning [47].

[5]https://www.iri.upc.edu/research/webprojects/barcelonarobotlab

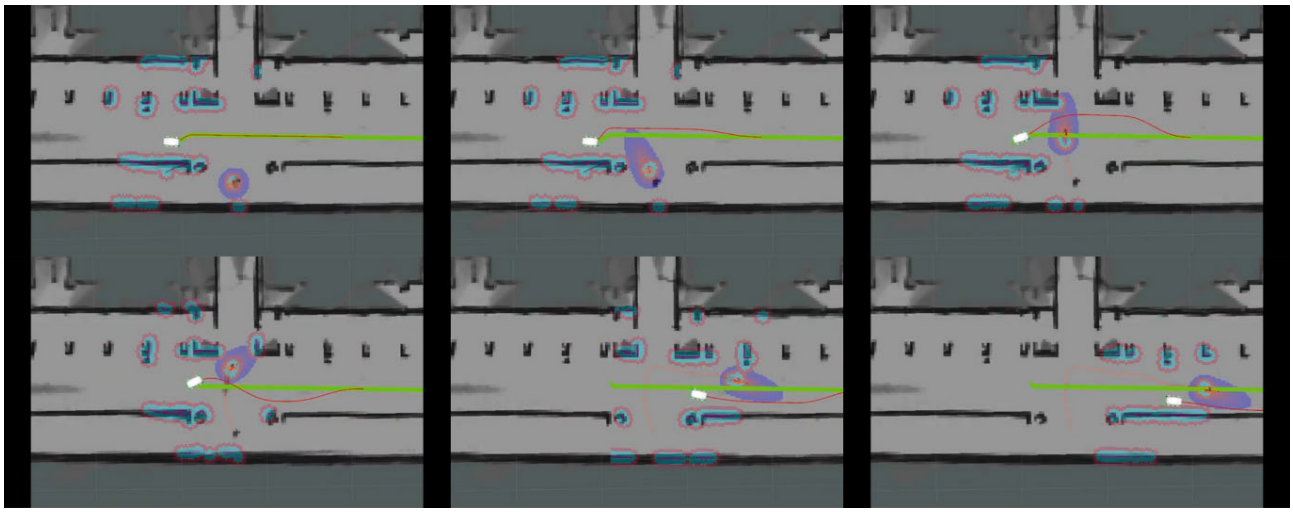


**FIGURE 5.** Path planning considering proxemic areas around a pedestrian.

Accurate pedestrian trajectory prediction is essential for improving the local planning capabilities of autonomous robots, ensuring safe navigation in dynamic urban environments. By accurately anticipating pedestrian movements, robots can proactively adjust their paths, minimizing the risk of collisions and enhancing the overall fluidity of movement within crowded spaces. Effective trajectory prediction directly contributes to smoother interactions between autonomous vehicles and pedestrians, facilitating safer coexistence in shared environments.

Collision avoidance strategies, usually integrated as reactive strategies, generally fall into four main categories: rule-based, force-field-based, model-based, and AI-based methods [2]. Rule-based methods utilize predefined rules and metrics, such as Time-To-Collision (TTC), to evaluate and prevent potential collisions. Force-field-based approaches, including potential field methods and elastic band techniques, apply attractive and repulsive forces to dynamically adjust robot paths and avoid obstacles [2]. Model-based strategies employ mathematical representations of pedestrian and vehicle dynamics to predict and navigate interactions, while AI-based methods utilize artificial intelligence techniques, such as neural networks and fuzzy controllers, to effectively handle complex and unpredictable pedestrian behaviors [64].

Trajectory prediction is commonly considered within the high-level path planning pipeline and can take advantage of similar families of methods, being more predominant those based on models and on learning [65]. In this case, model-based approaches can also incorporate physiological behaviors and environmental contexts to anticipate movements accurately. Learning-based methods, in contrast, are trained using historical pedestrian trajectory data to generate predictions of pedestrian paths, proving particularly advantageous in scenarios with complex or highly variable pedestrian behavior patterns [65], [66]. Furthermore, integrating social dimensions into trajectory prediction greatly enhances navigation efficiency and pedestrian comfort. Concepts like adaptive proxemics, which dynamically adjust robot-pedestrian distances based on comfort and social conventions, allow robots to navigate more naturally and respectfully in human-populated environments [62], [63], [67], [68]. This adaptive behavior, coupled with robust collision avoidance and accurate trajectory prediction, ensures more effective, socially acceptable, and reliable navigation outcomes.

In this work, we leverage a *proxemic layer* [69] to generate adaptive proxemic zones around pedestrians based on their predicted trajectories, which eventually falls into force-field methods category. By utilizing the capabilities of the PDT (based on the Spencer People Tracker), we provide real-time pedestrian poses (i.e., position, orientation and velocity in the *xyz* plane) as input to the proxemic layer node. This layer employs a Gaussian function to dynamically adapt the shape and size of the proxemic zones around each individual.

The behavior of the proxemic layer is formally defined by two zones. The more restrictive one, the *intimate zone*, is defined as the region where the Gaussian value exceeds a certain threshold $l$ (i.e., $g(x, y) > l$ in Eq. (1)). This region is treated as a critical obstacle, which the robot must never enter (i.e., we force a value close to 255 in the OGM). In contrast, the *personal zone* corresponds to values where $g(x, y) \leq l$ and is treated as a soft constraint, i.e., the robot attempts to avoid it but may traverse it if necessary.

$$h(x, y) = \begin{cases} g(x, y) & if g(x, y) \leq l \\ 253 & else \quad \text{(i.e., obstacle)} \end{cases} \tag{1}$$

where the *personal zone* is defined as

$$g(x, y) = e^{-(A(x-x_0)^2 + 2B(x-x_0)(y-y_0) + C(y-y_0)^2)}, \tag{2}$$

with

$$A = \frac{cos^2\Theta}{2\sigma_X^2} + \frac{sin^2\Theta}{2\sigma_Y^2},$$
$$B = \frac{sin2\Theta}{4\sigma_X^2} + \frac{sin2\Theta}{4\sigma_Y^2},$$
$$C = \frac{sin^2\Theta}{2\sigma_X^2} + \frac{cos^2\Theta}{2\sigma_Y^2}. \quad (3)$$

Here, $(x_0, y_0)$ represent the displacement of the Gaussian center with respect to the pedestrian's position. $\sigma$ represents the covariance and $\Theta$ is a rotation parameter allowing the orientation of the proxemic zones to align with individual-specific motion patterns.

Figure 4 shows the visualization of an experiment with the Ona robot and the proxemic layer in action. Here, a bystander is walking in the vicinity of the robot, which is detected and tracked (see the orange person track). In this Figure 4, we also show the colored local costmap, computed from the OGM's information. This costmap uses a pseudo-color code where red cells indicate an obstacle perimeter after applying a safety inflation radius (cyan). Purple regions are the proxemic area corresponding to the prediction of the pedestrian's motion. The local planner with an obstacle avoidance controller then computes a path avoiding this proxemic area (see the red path of the modified trajectory). Hence, Figure 5 illustrates how a pedestrian's proxemic zone can directly influence the robot's trajectory planning, requiring it to dynamically modify its planned path to safely circumvent the pedestrian and avoid interference.

Overall, the combined approach of precise pedestrian trajectory prediction and sophisticated collision avoidance techniques significantly improves local robot planning, facilitating safer, smoother, and socially compliant navigation in dynamic, pedestrian-rich environments. This example highlights the practical application of PDT outputs in shaping robot behavior but also illustrates ongoing challenges in robust prediction and real-time integration, especially in complex or cluttered urban environments. Such challenges align with those identified in recent literature [18], [48].

## V. CONCLUSION

In this work, we explored how the task of Pedestrian Detection and Tracking (PDT) can be approached based on the types of sensors mounted on a robot and the intended application of the PDT system. We present in a didactic manner the state of the art for various sensing modalities and categorize the observation models a robot can employ to perceive its environment regarding PDT. Finally, we discussed the key advantages that a PDT pipeline offers to autonomous robots navigating in crowded environments, supported by two examples of real-world experiments performed with our robot Ona in the Barcelona Robot Lab (north campus of the Universitat Politècnica de Catalunya, Barcelona).

This paper positions itself as a survey consolidating current knowledge on PDT methods and their integration into robotic navigation pipelines. The practical examples included serve as illustrative cases to contextualize the reviewed techniques, rather than as novel contributions.

Our study reaffirms the potential of using PDT frameworks within navigation pipelines to enhance robot performance in spaces shared with people. Potential applications include package delivery, search-and-rescue operations, and public transportation assistance, among others, as discussed throughout the literature.

In our examples, filtering the point cloud using pedestrian tracking qualitatively improved the clarity of SLAM-generated maps. Additionally, by predicting pedestrian trajectories and generating proxemic comfort zones, the robot adapted its path to avoid collisions and maintain socially acceptable distances. These outcomes are aligned with findings reported in related works but are presented here as qualitative demonstrations rather than benchmarked advancements.

Despite these positive results, several avenues for improvement remain. While the Spencer People Tracking system provides an efficient and lightweight solution, its performance can be affected by outliers. Integrating more robust clustering algorithms capable of distinguishing pedestrians directly at the clustering stage, rather than relying solely on overlapping detections, could increase reliability. Similarly, incorporating leg detectors for 2D laser scans may enhance the system's ability to differentiate between generic clusters and actual human detections. Extending the system with a panoramic, multi-camera setup that offers 360-degree coverage could also improve robustness by leveraging YOLO-based object detection across the full field of view.

In summary, this survey highlights the critical role of PDT for enabling safe, efficient, and socially-aware robot navigation, specially considering the imminent advent of mobile robots (e.g., humanoid, legged or wheeled platforms) being deployed in scenarios alongside pedestrians. While challenges remain in terms of robustness, scalability, and generalization, PDT continues to be a foundational component in the design of autonomous systems intended to operate in human-populated environments.

## REFERENCES

[1] R. Kümmerle, M. Ruhnke, B. Steder, C. Stachniss, and W. Burgard, "Autonomous robot navigation in highly populated pedestrian zones," *J. Field Robot.*, vol. 32, no. 4, pp. 565–589, Jun. 2015.

[2] T. Verstraete and N. Muhammad, "Pedestrian collision avoidance in autonomous vehicles: A review," *Computers*, vol. 13, no. 3, p. 78, Mar. 2024.

[3] S. Guillén-Ruiz, J. P. Bandera, A. Hidalgo-Paniagua, and A. Bandera, "Evolution of socially-aware robot navigation," *Electronics*, vol. 12, no. 7, p. 1570, Mar. 2023.

[4] P. Chong and Y. H. Tay, "A novel pedestrian detection and tracking with boosted HOG classifiers and Kalman filter," in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, Dec. 2016, pp. 1–5.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[6] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jun. 2007.

[7] Y. Ma, X. Chen, and G. Chen, "Pedestrian detection and tracking using HOG and oriented-LBP features," in *Network and Parallel Computing*, E. Altman and W. Shi, Eds., Berlin, Germany: Springer, 2011, pp. 176–184.

[8] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[9] R. Brehar, C. Vancea, and S. Nedevschi, "Pedestrian detection in infrared images using aggregated channel features," in *Proc. IEEE 10th Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2014, pp. 127–132.

[10] B. T. Bastian and C. V. Jiji, "Pedestrian detection using first- and second-order aggregate channel features," *Int. J. Multimedia Inf. Retr.*, vol. 8, no. 2, pp. 127–133, Jun. 2019.

[11] A. Verma, R. Hebbalaguppe, L. Vig, S. Kumar, and E. Hassan, "Pedestrian detection via mixture of CNN experts and thresholded aggregated channel features," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 555–563.

[12] D. A. Ribeiro, A. Mateus, P. Miraldo, and J. C. Nascimento, "A real-time deep learning pedestrian detector for robot navigation," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions (ICARSC)*, Jun. 2017, pp. 165–171.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[14] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[15] M. Razzok, A. Badri, I. El Mourabit, Y. Ruichek, and A. Sahel, "Pedestrian detection and tracking system based on deep-SORT, YOLOv5, and new data association metrics," *Information*, vol. 14, no. 4, p. 218, Apr. 2023.

[16] X. Xiao and X. Feng, "Multi-object pedestrian tracking using improved YOLOv8 and OC-SORT," *Sensors*, vol. 23, no. 20, p. 8439, Oct. 2023.

[17] Z. Zhou, W. Pei, X. Li, H. Wang, F. Zheng, and Z. He, "Saliency-associated object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9846–9855.

[18] L. Zhou, Z. Zhou, K. Mao, and Z. He, "Joint visual grounding and tracking with natural language specification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23151–23160.

[19] N. V. S. P. Nagulapati, S. R. Venati, V. Chandran, and R. Subramani, "Pedestrian detection and tracking through Kalman filtering," in *Proc. Int. Conf. Emerg. Smart Comput. Informat. (ESCI)*, Mar. 2022, pp. 1–6.

[20] D. R. Bruno and F. S. Osório, "Real-time pedestrian detection and tracking system using deep learning and Kalman filter: Applications on embedded systems in advanced driver assistance systems," in *Proc. Latin Amer. Robot. Symp. (LARS), Brazilian Symp. Robot. (SBR), Workshop Robot. Educ. (WRE)*, Oct. 2023, pp. 549–554.

[21] H. Wang, B. Wang, B. Liu, X. Meng, and G. Yang, "Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle," *Robot. Auto. Syst.*, vol. 88, pp. 71–78, Feb. 2017.

[22] J. Gómez, O. Aycard, and J. Baber, "Efficient detection and tracking of human using 3D LiDAR sensor," *Sensors*, vol. 23, no. 10, p. 4720, May 2023.

[23] K.-I. Na and B. Park, "Real-time 3D multi-pedestrian detection and tracking using 3D LiDAR point cloud for mobile robot," *ETRI J.*, vol. 45, no. 5, pp. 836–846, Oct. 2023.

[24] W. Wang, X. Chang, J. Yang, and G. Xu, "LiDAR-based dense pedestrian detection and tracking," *Appl. Sci.*, vol. 12, no. 4, p. 1799, Feb. 2022.

[25] X. Peng and J. Shan, "Detection and tracking of pedestrians using Doppler LiDAR," *Remote Sens.*, vol. 13, no. 15, p. 2952, Jul. 2021.

[26] J. Zhang, Z. Zhou, G. Lu, J. Tian, and W. Pei, "Robust 3D tracking with quality-aware shape completion," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 7160–7168.

[27] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, "Detecting and tracking people in real time with RGB-D camera," *Pattern Recognit. Lett.*, vol. 53, pp. 16–23, Feb. 2015.

[28] H. Liu, J. Luo, P. Wu, S. Xie, and H. Li, "People detection and tracking using RGB-D cameras for mobile robots," *Int. J. Adv. Robotic Syst.*, vol. 13, no. 5, Sep. 2016, Art. no. 1729881416657746.

[29] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 5636–5643.

[30] A. Saha, B. C. Dhara, S. Umer, K. Yurii, J. M. Alanazi, and A. A. AlZubi, "Efficient obstacle detection and tracking using RGB-D sensor data in dynamic environments for robotic applications," *Sensors*, vol. 22, no. 17, p. 6537, Aug. 2022.

[31] W. Liu, W. Li, T. Wang, J. He, and Y. Lou, "Real-time RGB-D pedestrian tracking for mobile robot," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, 2023, pp. 1–6.

[32] W.-Y. Hsu and P.-Y. Yang, "Pedestrian detection using multi-scale structure-enhanced super-resolution," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12312–12322, Nov. 2023.

[33] Y. Ma, X. Wu, G. Yu, Y. Xu, and Y. Wang, "Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery," *Sensors*, vol. 16, no. 4, p. 446, Mar. 2016.

[34] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vis. Image Understand.*, vol. 106, nos. 2–3, pp. 288–299, May 2007.

[35] F. Altay and S. Velipasalar, "The use of thermal cameras for pedestrian detection," *IEEE Sensors J.*, vol. 22, no. 12, pp. 11489–11498, Jun. 2022.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[37] C. Tian, Z. Zhou, Y. Huang, G. Li, and Z. He, "Cross-modality proposal-guided feature mining for unregistered RGB-thermal pedestrian detection," *IEEE Trans. Multimedia*, vol. 26, pp. 6449–6461, 2024.

[38] N. Bellotto and H. Hu, "Multisensor-based human detection and tracking for mobile service robots," *IEEE Trans. Syst., Man, Cybern., B (Cybernetics)*, vol. 39, no. 1, pp. 167–181, Feb. 2009.

[39] B. Musleh, F. García, J. Otamendi, J. M. Armingol, and A. De la Escalera, "Identifying and tracking pedestrians based on sensor fusion and motion stability predictions," *Sensors*, vol. 10, no. 9, pp. 8028–8053, Aug. 2010.

[40] Z. Zhao, X. Qi, Y. Zhao, J. Zhang, W. Wang, and X. Yang, "Pedestrian detection and tracking based on 2D LiDAR and RGB-D camera," in *Proc. 3rd Int. Conf. Control*, 2022, pp. 7–14.

[41] J. Liang, Y.-L. Qiao, T. Guan, and D. Manocha, "OF-VO: Efficient navigation among pedestrians using commodity sensors," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6148–6155, Oct. 2021.

[42] J. Lee, J.-S. Choi, E. Jeon, Y. Kim, T. Le, K. Shin, H. Lee, and K. Park, "Robust pedestrian detection by combining visible and thermal infrared cameras," *Sensors*, vol. 15, no. 5, pp. 10580–10615, May 2015.

[43] V. John, S. Tsuchizawa, Z. Liu, and S. Mita, "Fusion of thermal and visible cameras for the application of pedestrian detection," *Signal, Image Video Process.*, vol. 11, no. 3, pp. 517–524, Mar. 2017.

[44] Y. Xue, Z. Ju, Y. Li, and W. Zhang, "MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection," *Infr. Phys. Technol.*, vol. 118, Jul. 2021, Art. no. 103906.

[45] M. M. Islam, A. A. R. Newaz, and A. Karimoddini, "A pedestrian detection and tracking framework for autonomous cars: Efficient fusion of camera and LiDAR data," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 1287–1292.

[46] X. Dang, Z. Rong, and X. Liang, "Sensor fusion-based approach to eliminating moving objects for SLAM in dynamic environments," *Sensors*, vol. 21, no. 1, p. 230, Jan. 2021.

[47] M. Everett, Y. F. Chen, and J. P. How, "Motion planning among dynamic, decision-making agents with deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Jul. 2018, pp. 3052–3059.

[48] S. Li, L. Zhou, Z. Zhou, J. Chen, and Z. He, "MambaVLT: Time-evolving multimodal state space model for vision-language tracking," in *Proc. Comput. Vis. Pattern Recognit. Conf. (CVPR)*, Jun. 2025, pp. 8731–8741.

[49] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 797–803.

[50] E. Sisbot, A. Clodic, L. Marin U., M. Fontmarty, L. Brethes, and R. Alami, "Implementing a human-aware robot system," in *Proc. 15th IEEE Int. Symp. Robot Human Interact. Commun.*, Sep. 2006, pp. 727–732.

[51] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Apr. 2017, pp. 1343–1350.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[54] S. Eiffert, K. Li, M. Shan, S. Worrall, S. Sukkarieh, and E. Nebot, "Probabilistic crowd GAN: Multimodal pedestrian trajectory prediction using a graph vehicle-pedestrian attention network," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5026–5033, Oct. 2020.

[55] A. Santamaria-Navarro, S. Hernández, F. Herrero, A. López, I. del Pino, N. Rodríguez-Linares, C. Fernández, A. Baldó, C. Lemardelé, A. Garrell, J. Vallvé, H. Taher, A. M. Puig-Pey, L. Pagès, and A. Sanfeliu, "Toward the deployment of an autonomous last-mile delivery robot in urban areas: The ona prototype platform," *IEEE Robot. Autom. Mag.*, early access, Nov. 13, 2024, doi: 10.1109/MRA.2024.3487321.

[56] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2D LiDAR SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 1271–1278.

[57] I. D. Pino, A. Santamaria-Navarro, A. Garrell Zulueta, F. Torres, and J. Andrade-Cetto, "Probabilistic graph-based real-time ground segmentation for urban robotics," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 5, pp. 4989–5002, May 2024.

[58] T. Linder, S. Breuers, B. Leibe, and K. O. Arras, "On multi-modal people tracking from mobile platforms in very crowded and dynamic environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 5512–5519.

[59] T. Linder, F. Girrbach, and K. O. Arras, "Towards a robust people tracking framework for service robots in crowded, dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Apr. 2015, pp. 1–7.

[60] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2D range data," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3402–3407.

[61] A. Ortega, M. Silva, E. Teniente, R. Ferreira, A. Bernardino, J. Gaspar, and J. Andrade-Cetto, "Calibration of an outdoor distributed camera network with a 3D point cloud," *Sensors*, vol. 14, no. 8, pp. 13708–13729, Jul. 2014.

[62] A. Vega, L. J. Manso, P. Bustos, and P. Núñez, "A flexible and adaptive spatial density model for context-aware social mapping: Towards a more realistic social navigation," in *Proc. 15th Int. Conf. Control*, 2018, pp. 1727–1732.

[63] J. Ginés, F. Martín, F. J. Rodríguez-Lera, J. M. G. Hernández, and V. M. Olivera, "Defining adaptive proxemic zones for activity-aware navigation," in *Advances in Physical Agents II*, L. M. Bergasa, M. Ocaña, R. Barea, E. López-Guillén, and P. Revenga, Eds., Cham, Switzerland: Springer, 2021, pp. 3–17.

[64] B. Yang, S. Yan, Z. Wang, and K. Nakano, "Prediction based trajectory planning for safe interactions between autonomous vehicles and moving pedestrians in shared spaces," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 10513–10524, Oct. 2023.

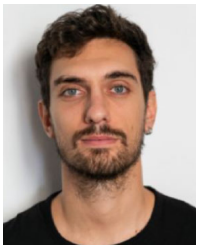
[65] M. Gulzar, Y. Muhammad, and N. Muhammad, "A survey on motion prediction of pedestrians and vehicles for autonomous driving," *IEEE Access*, vol. 9, pp. 137957–137969, 2021.

[66] X. Shi, X. Shao, Z. Guo, G. Wu, H. Zhang, and R. Shibasaki, "Pedestrian trajectory prediction in extremely crowded scenarios," *Sensors*, vol. 19, no. 5, p. 1223, Mar. 2019.

[67] C. Medina-Sánchez, S. Janzon, M. Zella, J. Capitán, and P. J. Marrón, "Human-aware navigation in crowded environments using adaptive proxemic area and group detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Apr. 2023, pp. 6741–6748.

[68] B. Bilen, H. Kivrak, P. Uluer, and H. Kose, "Social robot navigation with adaptive proxemics based on emotions," 2024, *arXiv:2401.17663*.

[69] T. Kirks, J. Jost, J. H. Finke, and S. Hoose, "Modelling proxemics for human-technology-interaction in decentralized social-robot-systems," in *Intelligent Human Systems Integration*, T. Ahram, W. Karwowski, A. Vergnano, F. Leali, and R. Taiar, Eds., Cham, Switzerland: Springer, 2020, pp. 153–158.



**N. PICELLO** received the B.Sc. degree in computer engineering and the M.Sc. degree in computer engineering, artificial intelligence and robotics from the University of Padova, in 2021 and 2024, respectively. He has been a Graduate Research Assistant with New York University and then a Research Technician with the Universitat Politècnica de Catalunya. His current research interests include the intersection between robotics and artificial intelligence, with a focus on computer vision, deep learning, and sensor fusion.

**F. HERRERO** received the B.Sc. degree in industrial engineering. He has been a Research Support Engineer with the Mobile Robotics Group, Institut de Robòtica i Informàtica Industrial, CSIC-UPC, a joint center of Spanish National Research Council (CSIC) and the Universitat Politècnica de Catalunya (UPC), Barcelona, since 2012. He has participated in various research projects related to robotics, navigation, human–robot interaction, and computer vision.

**S. HERNÁNDEZ** received the B.Sc. degree in telecommunications engineering and the B.Sc. degree in electronics engineering from UPC, in 2003 and 2005, respectively. He has been a Research Support Engineer with the Mobile Robotics Group, Institut de Robòtica i Informàtica Industrial (IRI), CSIC-UPC, since 2008. He has participated in various research projects related to robotics, navigation, human–robot interaction, and computer vision.

**A. LÓPEZ** received the B.Sc. degree in industrial engineering from the Technical University of Madrid (UPM), in 2015. He is currently a Robotics Research Technologist with the Institut de Robòtica i Informàtica Industrial (IRI), CSIC-UPC. He has participated in several research projects both in the public and private sectors. His work focuses on applied research in mobile robotics, human–robot interaction, and intelligent systems.

**A. SANTAMARIA-NAVARRO** received the M.S.Eng. and Ph.D. degrees from UPC, in 2012 and 2017, respectively. Prior to joining UPC, he was a Robotics Research Technologist with the NASA's Jet Propulsion Laboratory, USA. He is an Associate Professor with the Automatic Control Department, Universitat Politècnica de Catalunya (UPC). He is currently a PI of Spanish national and EU projects and his research interests include perception, cognition, and control of mobile robotic systems.