SCIENCE CHINA
Information Sciences

• **REVIEW** •

April 2025, Vol. 68, Iss. 4, 141101:1–141101:39
https://doi.org/10.1007/s11432-024-4303-8

Embodied navigation

Yunhao LIU[1*], Li LIU[1*], Yawen ZHENG[1*], Yunhuai LIU[2*], Fan DANG[1],
Ningbo LI[1] & Ke MA[1,3]

[1] *Tsinghua University, Beijing 100084, China*
[2] *Peking University, Beijing 100871, China*
[3] *Northwestern Polytechnical University, Xi'an 710072, China*

Received 20 August 2024/Revised 11 December 2024/Accepted 8 February 2025/Published online 13 March 2025

Abstract Navigation is a fundamental component of modern information application systems, ranging from military, transportations, and logistic, to explorations. Traditional navigations are based on an absolute coordination system that provides a precise map of the physical world, the locations of the moving objects, and the optimized navigation routes. In recent years, many new emerging applications have presented new demands for navigation, e.g., underwater/underground navigations where no GPS or other localizations are available, an un-explored area with no maps, and task-oriented navigations without specific routes. The advances in IoT and AI enable us to design new navigation paradigms, embodied navigation that allows the moving object to interact with the physical world to obtain the local map, localize the objects, and optimize the navigation routes accordingly. We make a systematic and comprehensive review of research in embodied navigation, encompassing key aspects on perceptions, navigation and efficiency optimization. Beyond advancements in these areas, we also examine the emerging tasks enabled by embodied navigation which require flexible mobility in diverse and evolving environments. Moreover, we identify the challenges associated with deploying embodied navigation systems in the real world and extend them to substantial areas. We aim for this article to provide valuable insights into this rapidly developing field, fostering future research to close existing gaps and advance the development of general-purpose autonomous systems grounded in embodied navigation.

Keywords artificial intelligence, Internet of Things, embodied AI, navigation, perception, LLM

Citation Liu Y H, Liu L, Zheng Y W, et al. Embodied navigation. Sci China Inf Sci, 2025, 68(4): 141101, https://doi.org/10.1007/s11432-024-4303-8

1 Introduction

Navigation has been a fundamental ability of autonomous systems, such as robots [1,2], AGVs (automated guided vehicles) [3–5], UAVs (unmanned aerial vehicles) [6–8], to fulfill a wide variety of tasks in complex environments [9–14], including housekeeping, package delivery, and disaster rescue. A navigation system locates itself, plans paths, and navigates toward target locations based on knowledge of the environment while avoiding potential obstacles. Such a complex process requires a deep integration of perception, decision-making, and motion control, all of which are essential to the successful deployment of such systems in real-world scenarios.

As the need for more intelligent, capable and versatile machines that can operate autonomously in diverse and complex environments has intensified, the scope and sophistication of navigation tasks have evolved accordingly. Navigation has developed from waypoint-based navigation in structured environments like factories and warehouses for basic automation to simultaneous localization and mapping (SLAM) with vision-based perception [15–17] for navigating in changing environments without predefined paths. With advancements in artificial intelligence (AI), human expectations for autonomous systems evolved from basic navigation to the execution of complex, goal-directed tasks. Modern navigation systems are now expected to perform high-level functions that go beyond mere movement. For example, autonomous vehicles must navigate not just for transportation but also adhere to traffic laws and ensure passenger safety; service robots are expected to autonomously navigate within homes and offices to assist with tasks like cleaning, organizing, or fetching objects. However, advancements in AI have

[*] Corresponding author (email: yunhao@tsinghua.edu.cn, liuli95@mail.tsinghua.edu.cn, yw-zheng21@mails.tsinghua.edu.cn, yunhuai.liu@pku.edu.cn)

been primarily driven by large-scale learning from human-curated, static internet data, which has led to significant improvements in tasks like image recognition, speech synthesis, and text understanding, sophisticated navigation tasks remain a formidable challenge for AI agents. Embodied navigation, with its emphasis on interaction with the physical world, represents a critical frontier in this pursuit. Unlike static data-driven navigation systems, embodied navigation systems must continuously perceive and respond to dynamic, often unpredictable environments under different task objectives, sometimes even human instructions [18].

We aim to provide a systematic and thorough review of research in embodied navigation. The literature on embodied navigation is categorized into four main aspects: perception, navigation, efficiency optimization, and embodied navigation enabled tasks.

- **Perception.** Perception lays the foundation for subsequent planning and control stages of navigation systems. In Section 2, we review perception models of embodied navigation systems to perceive and interpret their surroundings with diverse sensors and sensing modalities. It can be subdivided into geometric perception, which focuses on understanding the spatial layout and structure of environments, and semantic understanding, which involves recognizing objects and scenes within those environments while reasoning about their interrelationship and relationship to the navigation objective.
- **Navigation.** It includes the action planning and motion control (we focus more on the former in this article) process of the agents to navigate through complex environments, which is the core of navigation systems. Section 3 delves into various navigation methodologies that reflect the diverse strategies researchers employ to guide agents through physical environments to achieve specific goals, which mainly fall into two categories of approaches, geometry-based approaches that utilize traditional mapping and localization techniques, and learning-enhanced approaches that leverage advancements in machine learning to enhance navigational capabilities.
- **Efficiency optimization.** To meet the performance requirements of complex navigation tasks given the resource constraints that many navigation systems operate under, designated efficiency optimization for embodied navigation is essential. Section 4 addresses various strategies for latency optimization, energy efficiency optimization, and robustness improvement, which are essential for the practical deployment of these systems in real-world applications.
- **Embodied navigation enabled tasks.** With the advanced capability of embodied navigation, emerging tasks that require flexible mobility in diverse changing environments can be boosted. In Section 5, we review specific tasks enabled by embodied navigation, including autonomous driving, general assistant robots, navigation for bionic applications, and navigation in micro-environments. Each of these applications presents unique challenges and opportunities, highlighting the diverse potential of embodied navigation systems.

While embodied navigation holds great promise and has made significant strides in recent years, there remain substantial areas that require further improvement and in-depth study. At the end of this article, we also discuss these critical issues, highlighting the limitations and open questions that continue to hinder the full realization of embodied navigation's potential. Despite the advancements, challenges such as real-world applicability, multi-agent collaboration, bio-inspired neural architectures, and concerns around security and privacy still pose significant obstacles. These challenges must be addressed to propel the field forward and achieve the robust, versatile systems envisioned by researchers and practitioners alike.

2 Perception

In this section, we first present some theoretical foundations for embodied navigation in Subsection 2.1. On the basis of those foundations, the perception of the surrounding environment and the agent enables the following planning and control stages. In Subsection 2.2, we review the geometric perception task, whose main target is to reconstruct the geometric structure of the environment and/or localize the agent or the target via processing the data of one or multiple sensors. As a high-level understanding of the environment enables more complex navigation tasks such as object goal navigation and visual navigation, many recent efforts have been aimed at constructing a semantic or implicit representation of the environment. Furthermore, these hyper-semantic representations are superior in terms of robustness and storage efficiency. We regard obtaining these hyper-semantic representations and localizing the agent with such representations as hyper-semantic perceptions in Subsection 2.3.

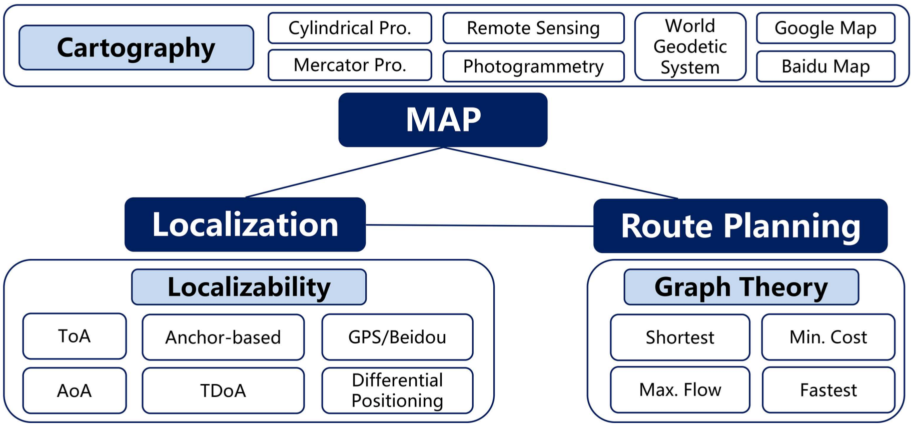

Figure 1 (Color online) Foundations and associated technologies for traditional navigation.

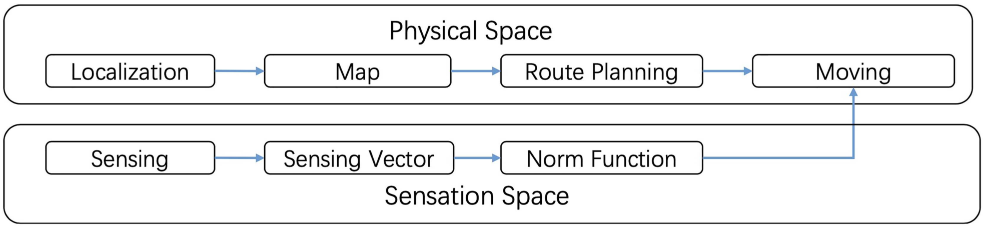

Figure 2 (Color online) Working process of an agent regarding sensation and physical space.

2.1 Theoretical foundations

Firstly, we provide a brief review of the theoretical foundations for traditional navigation. Subsequently, we introduce the sensation space which is spanned by the sensing data of embodied agents. We will show that the sensation space is a complete normed vector space, specifically a Banach space. It is consistent with the physical world, and $\boldsymbol{R}^3$ in terms of navigation. Additionally, some properties of the sensation spaces will be provided.

2.1.1 *Foundations for traditional navigation*

Traditional navigation is based on three theoretical foundations and the associated technologies. As depicted in Figure 1, cartography contributes to the creation of maps, employing cylindrical projections, Mercator projection, photogrammetry, remote sensing, the world's measurement system, and other geographic information systems (GIS). Localization techniques utilize triangulation to determine the position of an embodied agent on the map. Examples of localizations include the navigation satellite system (GPS, Beidou, GNSS, etc.), differential global positioning system (DGPS), landmarks with anchor nodes, triangulations based on time of arrival (ToA), time difference of arrivals (TDoA), and angle of arrival (AoA). With the map and the real-time location of the agent on the map, graph theory is applied to guide the agent's movement routes. Various optimization goals are employed to generate different routes, such as the shortest path, maximum flow, minimum cost on routes, and the fastest routes.

2.1.2 *Sensation space (vector space)*

During movement, a moving agent acquires environmental information via sensing capabilities like satellite signals or wireless signal strength. It obtains a map, often constructed in prior, localizes itself on the map, and employs route optimization algorithms to guide its real-time movement. This working process is illustrated in Figure 2. Based on years of practical experience, this approach is sufficient and effective, while an open question remains. Is this the only necessary way to navigate?

We can observe that sensing data essentially guides the movement of the agent, rather than the map or routing algorithm determined prior to navigation. Sensing data alone, without being mapped to the physical world, may be sufficient to guide the movement. To answer this question well, we have the following observations.

Definition 1 (Sensation space). Suppose the moving agent is equipped with certain sensing capabilities, providing sensing data continuously. The sensing data are multi-modality coming from various sources such as the wireless signals (signal strength, amplitudes of the signals, phase of the signals, TDoA), accelerometers, gyroscope, and camera (images and the parameters). Without loss of the generality, we denote these sensing data as a data vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$, and have the following theorem.

Theorem 1 (The sensing data $\boldsymbol{s}$ form a vector space $\boldsymbol{S}$). To show that the sensing data $\boldsymbol{s}$ form a vector space $\boldsymbol{S}$, we have the following statements.

- Commutative law. For all sensing vectors $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{S}$, $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$.
- Associative law. For all vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{S}$, $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$.
- Additive identity. For any vector $\boldsymbol{x} \in \boldsymbol{S}$, the vector space contains the additive identity element $\boldsymbol{0}$ such that $\boldsymbol{0} + \boldsymbol{x} = \boldsymbol{x} + \boldsymbol{0} = \boldsymbol{x}$.
- Additive inverse. For each vector $\boldsymbol{x} \in \boldsymbol{S}$, there is an $-\boldsymbol{x} \in \boldsymbol{S}, -\boldsymbol{x} + \boldsymbol{x} = \boldsymbol{0}$.
- Distributivity. $c(\boldsymbol{x} + \boldsymbol{y}) = c\boldsymbol{x} + c\boldsymbol{y}$ and $(\boldsymbol{x} + \boldsymbol{y})c = c\boldsymbol{x} + c\boldsymbol{y}$.
- Scalar associativity. $c(d\boldsymbol{x}) = cd\boldsymbol{x}$.
- Identity. $1\boldsymbol{x} = \boldsymbol{x}$.

Notice that in order to keep the physical meaning of the sensing data in the physical world, the additional operations in vector space $\boldsymbol{V}$ may not be the one in the physical world $\boldsymbol{R}$. For instance, sensing data from images shall be converted to the frequency space to operate. A red plus a blue may generate a purple, which violates the closeness of additional operations. Thus the red and blue shall all be converted to the corresponding frequency and the strength, added in the frequency space, and then converted back to the purple in the color space.

Norm in sensation space. In order to navigate in the sensation space, we need to define the norm in the sensation space $\boldsymbol{S}$. This norm shall be consistent with the physical space so that the navigation in two spaces is consistent as well. In practice, since sensing vectors are obtained in the physical world, every sensing vector $\boldsymbol{x} \in \boldsymbol{S}$ is associated with one physical location $\boldsymbol{r} \in \boldsymbol{R}$ only. Let $f\colon \boldsymbol{S} \to \boldsymbol{R}$ be the map function from the sensation space $\boldsymbol{S}$ to the physical world space $\boldsymbol{R}$. We define the norm function D in $\boldsymbol{S}$ as follows:

$$\forall \boldsymbol{x} \in \boldsymbol{S},\ D(\boldsymbol{x}) = |f(x)|_2, \tag{1}$$

where $|\cdot|_2$ is an l_2 norm. By theorem in vector spaces, $\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{S}, d(\boldsymbol{x}, \boldsymbol{y}) = |D(\boldsymbol{x}) - D(\boldsymbol{y})|$ defines a metric on $\boldsymbol{S}$. And this metric is consistent with Euclidean distance in $\boldsymbol{R}$.

2.1.3 *Isometry of sensation subspaces*

The sensation space is a Banach space, i.e., a complete vector space. As its cardinality is much larger than the physical space $\boldsymbol{R}^3$, we have the following corollary by the theorems of vector spaces.

- There exists a sensation subspace, which is isometric with $\boldsymbol{R}^3$. The key to finding such subspace is to define a distance-preserving map function f such that for any $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{S}$, $d_{\boldsymbol{S}}(\boldsymbol{x}, \boldsymbol{y}) = d_{\boldsymbol{R}^3}(f(\boldsymbol{x}), f(\boldsymbol{y}))$.
- The distance-preserving map function f is injective, otherwise two distinct points, say $\boldsymbol{a}$ and $\boldsymbol{b}$, could be mapped to the same point, which contradicts the coincidence axiom of the metric, i.e., $d(\boldsymbol{a}, \boldsymbol{b}) = 0$ iff $\boldsymbol{a} = \boldsymbol{b}$. Moreover, an order embedding between partially ordered sets is injective as well. Clearly, the isometry between the sensation subspace and the physical world is a topological embedding.
- In practice, every point in the physical space, if it can be used during navigation, will have the corresponding sensation data. Thereby the isometry function f is bijective, i.e., global isometry, and has the function inverse.
- Every time series $\{r(t)\} \in \boldsymbol{R}$ has the corresponding $\{s(t) = f(r(t))\} \in \boldsymbol{S}$ such that $d(r(t_1), r(t_2)) = d(f(r(t_1)), f(r(t_2)))$.

2.2 Geometric perception

Geometric perception aims to reconstruct the geometric 3D/2D structure of the environment and obtain the geometric position of the agent or the navigation target. As shown in Figure 3, geometric perception can be achieved using different sensors, and there exist four geometric perception mechanisms with distinct outputs. The first mechanism is SLAM (see Subsection 2.2.1), where the agent tries to simultaneously construct a map of the environment and localize itself in the map in real-time. SLAM enables an embodied agent to navigate through an unknown environment autonomously. The second is the structure from motion (SfM) (see Subsection 2.2.2), where the goal is to reconstruct the geometric structure of a

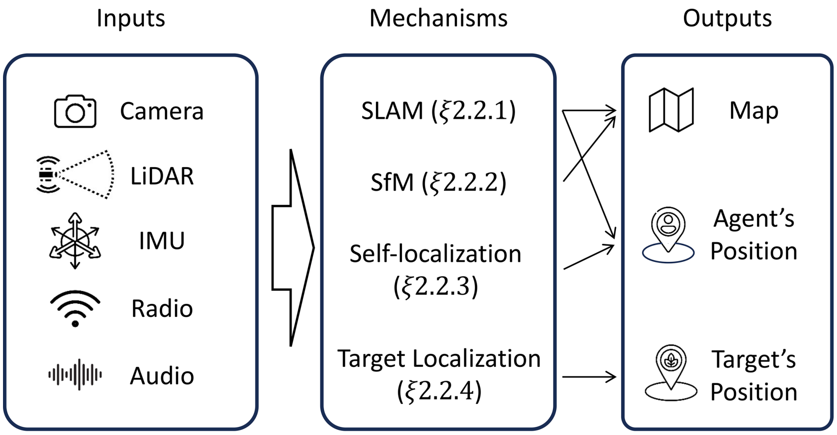

Figure 3 Geometric perception.

large-scale environment based on an unordered set of frames captured by one or many agents. Compared with SLAM, SfM can achieve a higher reconstruction accuracy and supports the reconstruction of a larger environment with its looser requirement on the reconstruction duration. The third is self-localization (see Subsection 2.2.3), where the agent aims to localize itself in a known environment whose geometric map is constructed by the SLAM program on other agents or SfM in advance. The last is target localization (see Subsection 2.2.4), where the agent aims to localize the navigation target.

2.2.1 *SLAM*

SLAM refers to the process of simultaneously constructing a geometric map of the environment (i.e., mapping) and localizing the agent in the autonomously generated map (i.e., localization). It allows camera-equipped agents to automatically understand an unknown environment without relying on any manually made map or pre-deployed infrastructure and in real-time. Odometry, loop closure detection, and backend are the three most important sub-processes in SLAM. Odometry focuses on local mapping and localization by analyzing a sequence of inputs; loop closure detects whether the agent reaches a previously visited place; and the backend module further optimizes the results of the front-end modules (i.e., odometry and loop closure detection) towards the goal of maximizing the posterior probability.

Process I. Odometry. We classify different odometry methods based on their underlying sensor modalities.

- **Visual odometry.** In this part, we first introduce the two classical visual odometry methods, i.e., the feature method and the direct method. As the classical methods have several challenges, many deep learning-based methods have been proposed to address the challenges. We also systematically review these deep learning-based techniques.

The feature method workflow is illustrated in Figure 4. First, features like SIFT [19], SURF [20], and ORB [21] are extracted from each frame. Features are typically corners detectable in multiple images, consisting of a keypoint (position) and a descriptor (surrounding image information). In monocular visual SLAM, keypoints are 2D coordinates, while in stereo or RGB-D SLAM, they can be 3D points with depth or 2D points if depth sensing fails [22]. Keypoints are then matched across frames or with 3D landmarks on the map using feature descriptors. Camera poses are estimated by minimizing reprojection errors of matched keypoints, assuming they correspond to static objects. Common methods include epipolar geometry (2D-2D matches), iterative closest point (ICP) (3D-3D matches), and PnP (3D-2D matches). Finally, 2D keypoints' depths are determined through methods like triangulation to create new 3D landmarks on the map.

Another commonly used visual odometry method is the direct method. Instead of matching the extracted features, the direct method directly estimates the relative pose among different frames by minimizing the photometric error. Despite its advantage in efficiency, robustness to featureless scenarios, and the ability to construct dense or semi-dense geometric maps, the direct method heavily relies on the photometric invariance assumption and may be more vulnerable to scenarios of varying or complex illumination than the feature method.

Although the two methods can work well in many scenarios, they suffer from caused by weak robustness

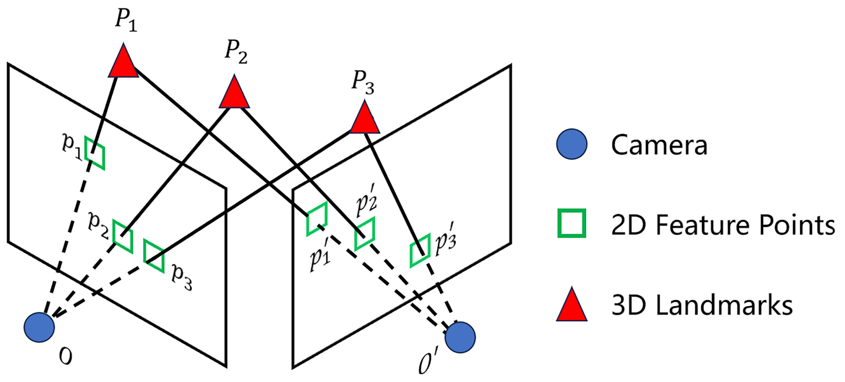

Figure 4 (Color online) Example of the epipolar geometry-based feature method. The camera takes two frames successively and its pose changes from O to O'. First, a total of six feature points (i.e., p_1, p_2, p_3, p_1', p_2', p_3') are detected from the two frames. Using the feature descriptor, three pairs of matched features are generated, including (p_1, p_1'), (p_2, p_2'), and (p_3, p_3'). The epipolar geometry assumes that each pair corresponds to a static 3D landmark whose coordinate is unknown. The relative camera pose between O' and O is obtained under such an assumption. Finally, the coordinates of the 3D landmarks (P_1, P_2, and P_3) are obtained via triangulation.

to dynamic scenes and lighting/weather conditions. Traditional methods [15,23–31] have been proposed to work around the issue by optimizing brightness/photometric info. However, the robustness improvement from these hand-crafted strategies is relatively limited. To meet these challenges, many deep-learning methods have been developed. To handle dynamic scenes, researchers propose to identify non-rigid contexts via semantic segmentation and then cull points of non-rigid contexts from the feature or pixel set used by visual odometry. An early effort on non-rigid context culling (NRCC) is Pair-Navi [32], a peer-to-peer indoor navigation system based on trajectory sharing between leader and followers in visual SLAM. A later system known as EdgeSLAM [33] enables the process running on resource-constrained mobile devices by strategically offloading the NRCC task to powerful edge servers. For the challenge of dynamic lighting/weather conditions, recent studies propose to replace the original intensity values or handcrafted descriptors with deep feature descriptors [34, 35].

• **LiDAR odometry.** Deep learning-based visual SLAM has made progress, but light detection and ranging (LiDAR)-based SLAM is popular in industry due to its better performance in challenging conditions like fog and varying light [16]. LiDAR creates point clouds, which are collections of points with precise angle and distance data. It works by sending out laser beams and measuring their return time [17]. LiDAR odometry calculates movement by matching features in point-cloud maps from different times. LiDAR systems come in 2D and 3D types: 2D is used for simple indoor spaces, while 3D is better for complex outdoor areas with more spatial information.

The key to processing LiDAR odometry is scan matching, which determines the position of an agent by analyzing consecutive LiDAR scans of point-cloud maps. Similar to visual odometry, LiDAR odometry can be categorized into (i) direct methods, (ii) feature-based methods, and (iii) deep learning-based methods, depending on how scan matching is performed [36].

The direct methods are a kind of method employed to estimate motion by directly comparing point cloud data obtained from LiDAR scans. Unlike other matching methods, such as feature methods, direct methods eschew the extraction of specific features from the point cloud. Instead, it leverages the entirety of the point cloud data for alignment. The primary objective of these methods is to estimate the sensor's motion trajectory by minimizing the geometric discrepancies between consecutive point clouds [37]. There can be two mainstream direct methods for LiDAR odometry: ICP and normal distributions transform (NDT). ICP iteratively matches points in one cloud to the nearest points in another, updating the transformation to minimize distances between corresponding points until convergence [38]. It is simple and effective, though sensitive to initial alignment and noise.

Feature-based methods in LiDAR odometry extract and match distinct features from point clouds to estimate motion. They focus on specific geometric structures like edges, corners, or planes, improving efficiency and accuracy compared to direct methods that use the entire point cloud. These features are matched across consecutive scans to determine relative motion. LOAM [39, 40] is a popular method that extracts edge and planar features, matching them to estimate motion and build a map. It balances real-time performance and accuracy. LeGO-LOAM [41], an extension of LOAM for ground vehicles, optimizes feature extraction and matching, targeting flat ground surfaces for efficiency. R-LOAM [42] further extends LOAM by incorporating prior knowledge of a reference object to enhance localization accuracy. It uses a known 3D model and its global position, adding mesh features to point features for optimization. R-LOAM employs an axis-aligned bounding box tree for efficient mesh feature matching, reducing absolute pose error and improving map quality.

Deep learning methods for LiDAR odometry are becoming more popular as they can improve accuracy and efficiency in motion estimation and mapping [43–46]. LO-Net [47] is a real-time deep learning framework that uses a convolutional network to learn features and capture data dynamics. It includes a special loss function and a scan-to-map module to enhance accuracy. Another approach by Liu et al. [48] used a bird's eye view projection and deep learning for accurate motion estimation. Their method addresses common challenges and outperforms traditional SLAM methods on the KITTI dataset, achieving low drift over long distances.

- **Radio odometry.** Radio odometry utilizes radio signals to estimate position and movement. It employs various radio technologies such as WiFi, ultra-wideband (UWB), and millimeter wave (mmWave) to determine the agent's location and trajectory. By measuring and analyzing the strength, delay, ToA, and relative position of radio signals, radio odometry can provide accurate positioning and navigation information [49, 50].

WiFi-based SLAM is valuable for research and daily life due to its widespread coverage. Liu et al. [51] presented C-SLAM-RF, a system using WiFi RSS and smartphone-based PDR for large indoor environments. Arun et al. [52] introduced P2SLAM, which uses WiFi for indoor SLAM to address issues with other sensors in certain environments. It extracts features from WiFi signals and integrates them with odometry. Recently, WAIS [53] was proposed, combining WiFi sensing with Visual-SLAM to reduce computational and memory needs. WAIS improves accuracy while significantly reducing resource usage, and provides an open-source toolbox called WiROS for WiFi measurements.

UWB and mmWave technologies offer higher precision localization due to their shorter wavelengths. Wang et al. [54] combined UWB with visual-inertial odometry to create drift-free maps in real-time, even in challenging environments. Cao et al. [55] used multiple UWB nodes for relative localization between robots, improving accuracy through optimization. For mmWave, Palacios et al. [56] proposed CLAM, a distributed SLAM algorithm for 5G networks that addresses beam-training and device association challenges. He et al. [57] presented a low-cost mmWave SLAM scheme that achieves submeter-level accuracy without extensive hardware requirements. These approaches demonstrate how UWB and mmWave can enhance localization and mapping in various scenarios.

Radio offers wide applicability and low power use, making it promising for odometry [50]. Researchers are exploring various radio technologies like Bluetooth [4], radio frequency identification (RFID) [58], and Zigbee [59], or combining them [60] for better localization and motion estimation. GPS is usually combined with other sensors rather than used alone for odometry [61, 62]. When used with inertial measurement units (IMUs), cameras, LiDAR, or other radio systems, GPS fusion can improve odometry and navigation accuracy.

- **Inertial odometry.** Inertial odometry is a method of estimating the position and orientation changes of agents using data from an IMU. By integrating the acceleration and angular velocity data provided by the IMU, the agent's velocity and position can be computed. Compared to visual odometry and LiDAR odometry, inertial odometry is commonly used in environments where external signals are unreliable or unavailable, such as underground, indoors, or other GPS-denied environments.

TLIO [63] proposes an extended Kalman filter (EKF) framework for IMU-only state estimation, integrating relative state estimates from IMU measurements with a neural network that regresses 3D displacement estimates and their uncertainties. This method addresses the significant drift caused by sensor bias and noise, outperforming traditional velocity integration and AHRS filters in position and orientation estimation. Chen et al. [64] used deep recurrent neural networks to estimate user displacement from raw inertial measurements, formulated as an optimization problem. It addresses error growth and dependency on fixed sensor positions or periodic motion patterns. The method demonstrates high accuracy in diverse conditions, including non-periodic motions like shopping trolley tracking and dynamic activities like running. DUET [65] is a deep learning-based IMU online calibration method designed to compensate for runtime errors in accelerometers and gyroscopes, improving inertial-based odometry. By using a differential error learning strategy, it learns sensor errors from displacement and orientation increments, mitigating integration errors during odometry computation. Experiments on public visual-inertial datasets show a 20% improvement in position estimation accuracy, comparable to existing methods but with lower complexity.

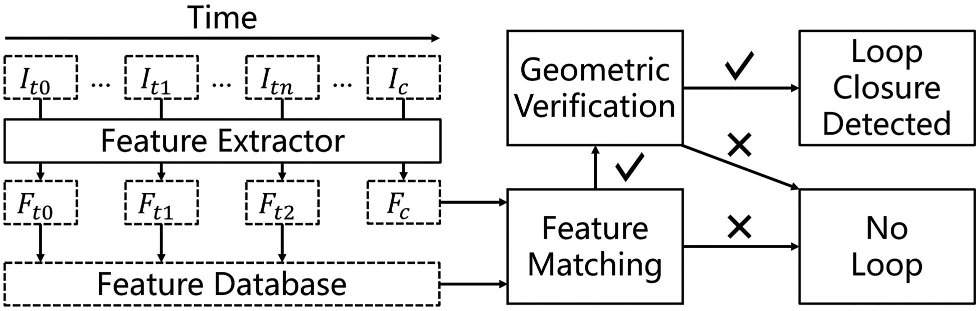

Figure 5 Typical workflow of visual loop closure detection.

As previously mentioned, inertial odometry provides high-frequency data updates, allowing it to quickly respond to dynamic changes in extreme environments [66]. However, due to the nature of integration, unbounded error drifts in inertial odometry accumulate over time, posing a major challenge. Therefore, IMUs typically play a supplementary role in odometry, often combined with other methods like visual odometry [67] and LiDAR odometry [68]. By providing high-frequency motion information (e.g., angular velocity and acceleration), IMUs compensate for the shortcomings of other sensor data, enhancing the accuracy and robustness of the system's localization and mapping. These studies will be elaborated on later.

• **Odometry accross different modalities.** Visual odometry excels in detailed environment mapping and loop closure detection, LiDAR odometry provides high-precision measurements under adverse illumination conditions, radio odometry offers wide applicability and low power consumption, and inertial odometry delivers high-frequency updates. Combining these methods to establish multimodal odometry enhances the overall system's robustness and accuracy, compensating for the individual limitations of each sensor type [69–72].

Combining visual and LiDAR odometry is a natural progression in the field [73–75]. Zhang et al. [76] used visual odometry for initial motion estimation and LiDAR for refinement, improving robustness and accuracy in challenging conditions. Environmental features like lines [77] and planes [78] have become more important, enhancing vision-LiDAR fusion in complex environments. Recently, NALO-VOM [5], a LiDAR-guided monocular visual odometry system for UGV navigation, addresses sparse environment maps in traditional methods. It uses a plane prediction network trained on LiDAR data to achieve scale-consistent camera poses and a semi-dense map, improving navigation capabilities.

IMUs are often combined with other sensors for odometry, providing high-frequency motion data for real-time calibration. In visual-inertial odometry [79, 80], ESVIO [81] introduces event-based stereo visual-inertial odometry, using event streams, standard images, and inertial measurements. It includes ESIO for event-based processing and ESVIO for integrating image-aided event streams, improving state estimation in challenging environments. For LiDAR-inertial odometry [82, 83], AdaLIO [84] tackles parameter degeneracy in narrow spaces by dynamically adjusting voxelization and normal vector estimation based on IMU-detected environment types. This approach improves odometry in confined spaces, as shown in public datasets. LiDAR, visual, and inertial odometry fusion combines the strengths of all three sensors [85, 86], addressing issues often found in LiDAR-inertial methods [72]. LVIO-Fusion [87] integrates LiDAR-visual-inertial odometry and mapping for robust state estimation and precise mapping in challenging environments. It combines dynamic voxel mapping LiDAR-inertial odometry with direct image projection for optical flow matching, and visual-inertial odometry with coarse-to-fine state estimation. Evaluations demonstrate its superior accuracy and robustness in challenging scenarios.

Process II. Loop closure detection. When the agent revisits a previously visited place, it should be able to detect the loop ideally, with which it can optimize the estimated pose and the reconstructed 3D environment. However, as odometry aims to estimate the relative pose, the estimation drift will accumulate over time and may lead to an unrecognized loop. By employing the loop closure detection technique to identify the loop, the loop can be closed and lead to a more accurate perception of the geometric structure as well as the agent's pose.

• **Vision-based methods.** The common workflow of vision-based (or appearance-based) loop closure detection is shown in Figure 5. The camera continuously takes images of the surrounding environment. Keyframes (i.e., $I_{t0}, I_{t1}, \ldots, I_{tn}$) are then sampled from the stream of images, using sampling strategies such as uniform sampling in time [88], uniform sampling in space [89], and uniform sampling in appearance [90]. The feature descriptors [91–93] of these keyframes are extracted and then stored in the feature database. The same feature extractor is employed for the current frame I_c, and the generated feature descriptor F_c is matched with those in the feature database. If F_c is very similar to some feature descriptor in the database, a more stringent geometric verification may be further employed to avoid false positive detections.

• **LiDAR-based methods.** As demonstrated in odometry, it is noted that compared to visual sensors (e.g., RGB cameras), LiDAR provides more robust perceptual information under varying illumination conditions. However, in loop closure detection, LiDAR point clouds only contain geometric information and lack the rich information present in images that is essential for place recognition [94]. Consequently, LiDAR-based methods are relatively rare in current research. The workflow of LiDAR-based loop closure detection is similar to that of vision-based methods, with feature descriptors for LiDAR-based methods primarily including histogram-based descriptors and segmentation-based descriptors. Histogram-based descriptors represent point cloud maps as histograms, such as a set of normal distribution [95, 96] and 3D Gestalt descriptors [97]. Segmentation-based descriptors calculate object information as a preprocessing step, representing point cloud maps as semantic objects [98–100], which significantly reduces matching time. Recently, PADLoC [101] is proposed as a novel method for joint loop closure detection and registration in LiDAR-based SLAM, using a transformer-based head for point cloud matching and a unique loss that reframes the matching problem as a classification task for the semantic labels.

• **Radio-based methods.** As radar sensors benefit from widespread deployment and the high-precision mapping capabilities, radio-based loop closure detection, though less studied than vision-based and LiDAR-based methods, has garnered increasing attention in recent years [102]. Compared to images, radar images have less distinctive pixel-level features for descriptors and are affected by multi-path reflection problems, making radio-based loop closure detection more challenging. RadarSLAM [103] utilizes M2DP [37], a rotation-invariant global descriptor, to represent FMCW radar images initially converted to point clouds. In M2DP, the left and right singular vectors of the density signatures of the point cloud are used as descriptors. To reduce cumulative errors in radio-based loop closure detection for multi-path SLAM, Gao et al. [104] proposed Wi-Loop SLAM, including a delayed selection strategy and a Bayesian model with a particle-based sum-product algorithm (SPA) to ensure accurate loop optimization, effectively correcting state estimate drift in environments with obscured propagation paths. For robust SLAM in large-scale environments, TBV [105] introspectively verifies loop closure candidates by combining multiple place-recognition techniques [106, 107] and delaying loop selection until after verification, addressing the challenge of false constraints and integrates this with a robust odometry pipeline.

• **Loop closure detection across different modalities.** Like odometry, multi-sensor fusion methods can combine the advantages of different modalities to achieve more robust loop closure detection performance [108–111]. VINS-Mono [112] utilizes DBoW2 [113], a state-of-the-art bag-of-words place recognition approach, for loop detection with image frames and uses IMUs to achieve high-frequency state estimation for closed-loop closure. For LiDAR-inertial systems, Shan et al. [114] implemented Euclidean distance-based loop closure detection on point cloud maps, where IMUs are used to estimate the agent's motion to de-skew point clouds and serve as an initial guess for LiDAR odometry optimization. Then, LVI-SAM [85], a framework for tightly-coupled lidar-visual-inertial odometry, proposes a two-step method where loop closures are first identified by the visual-inertial system and further refined by the LiDAR-inertial system. Recently, Chghaf et al. [115] presented a novel approach to loop closure detection in SLAM by leveraging multiple modalities through similarity-guided particle filtering (SGPF) for the search, integrating bag-of-words for camera-based and scan context for LiDAR-based place recognition. TS-LCD [116] is introduced as a multi-sensor fusion-based loop-closure detection scheme, employing a timestamp synchronization method based on data processing and interpolation, and a two-order loop-closure detection scheme for the fusion validation of visual and laser loops.

Process III. Backend. In essence, odometry and loop closure detection solve the problem of short-term and long-term data association, respectively. As data from different modalities have distinct characteristics, different methods are proposed to handle data association for different modalities. Despite the significant differences in odometry and loop closure detection methods used for different modalities, the back-end optimization techniques employed are highly consistent across these modalities. The reason is that the backend simply utilizes the data association relationship provided by the front-end modules to construct the observation equation (detailed below).

Formally speaking, the goal of the backend is to estimate the values of the unknown x_k (the state of the agent and the coordinates of observed landmarks at time step k) given z_k (the observation data related to landmarks) and u_k (the data of a motion-related sensor such as an IMU or a wheel encoder). Two fundamental equations are commonly used to describe the state estimation process. One is the motion equation

$$x_k = f(x_{k-1}, u_k) + w_k \tag{2}$$

with f abstracting the motion process and $w_k \sim \mathcal{N}(0, R_k)$ denoting the random noise. The other is the observation equation

$$z_k = h(x_k) + v_k \tag{3}$$

with h abstracting the observation process and $v_k \sim \mathcal{N}(0, Q_k)$ denoting the random noise. The correspondence between z_k and x_k is provided by the modality-specific front-end module. The common optimization goal is maximum a posteriori (MAP), i.e., maximizing the probability $P(x|z, u)$. With reasonable approximations, the MAP problem can be refactored as a least square problem

$$\underset{x,y}{\operatorname{argmin}} \left(\sum_k ||f(x_{k-1}, u_k) - x_k||^2_{R_k} + \sum_{k,j} ||h(y_j, x_k) - z_{k,j}||^2_{Q_{k,j}} \right). \tag{4}$$

A common solution to the problem is the Kalman filter [117]. The Kalman filter performs estimation in an incremental manner. At each time step k, the Kalman filter estimates the value of x_k and does not aim to update the previous unknown variables (i.e., $x_0, \ldots, x_{k-1}$) with the latest data (i.e., z_k and u_k).

The estimation at time step k consists of two steps. The first step predicts the a priori state estimation $\hat{x}_k^-$ following the motion equation (2). The second step corrects the a priori state estimation $\hat{x}_k^-$ to obtain the final estimation result — the a posteriori state estimation $\hat{X}_k$. The correction is achieved via

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - h(x_k)). \tag{5}$$

Intuitively, $z_k - h(x_k)$ acts as a feedback signal reflecting the difference between the predicted observation $h(x_k)$ and the actual observation z_k.

However, the Kalman filter is initially designed for the simple case of fixed and linear motion and observation processes as well as known Gaussian noises, so it does not lead to optimal results in real-world SLAM systems of more complexity. Many variants of the Kalman filter are proposed to handle the challenge and a relevant survey can be found in [118].

Another mainstream solution to the backend is graph-based optimization. In graph-based optimization, vertices represent unknown variables while edges represent motion and observation equations. The optimization goal of graph-based optimization is still an MAP or equivalently a least square problem, but graph-based optimization shares several differences with the Kalfam filter or its variants. First, graph-based optimization typically solves the problem via non-linear optimization methods such as the Gauss-Newton method and the Levenberg-Marquardt method, which are more memory-efficient and thus more suitable for large-scale environments than the Kalman filter or its variants. Second, graph-based optimization can update an unknown variable based on its past, current, and future equations, while the Kalman filter or its variants operates incrementally, i.e., only updating the estimation of the current pose based on the current motion and observation. Consequently, graph-based optimization can achieve a higher accuracy. Due to the above advantages, graph-based optimization has gained more popularity recently and has been integrated into several well-known SLAM systems, such as maplab [119, 120] and ORB-SLAM3 [121].

2.2.2 *SfM*

SfM is somehow similar to SLAM — it also tries to reconstruct the 3D environment from a series of sensor data. However, the two concepts are not the same as they share some differences. First, SfM puts more emphasis on reconstruction than on localization and puts more emphasis on reconstruction accuracy rather than reconstruction throughput. Second, due to the looser requirements on reconstruction throughput, SfM approaches can be used to reconstruct an extremely large environment. For example, a worldwide 3D SfM point cloud has been constructed by processing 9.2 billion panoramic images from Google street view [122], and a world-scaled SfM-based modeling is achieved by operating on the Yahoo

100 million image dataset [123]. Third, the input data of the SfM problem is more irregular — the data may be captured by different sensor models and their capturing time may be unordered. For example, the aforementioned Yahoo 100 million image dataset is created by collecting user-uploaded images and videos from Yahoo's Flickr image and video sharing platform [124]. In contrast, SLAM involves a set of ordered sensor data from a fixed sensor (or a fixed set of sensors in the case of multi-modal SLAM).

Traditional applications of SfM include geosciences and cultural heritage documentation. As more and more camera-equipped autonomous vehicles are appearing on the road, researchers have proposed to use SfM to reconstruct urban scenes offline from crowdsourced images recently [125], and the reconstructed scenes, in turn, can guide navigation of these autonomous vehicles online. The impact of various camera setups, weather conditions, and algorithms on the reconstruction quality is investigated. Several useful guidelines regarding the data collection are further proposed based on the investigations aforementioned.

With the prevalence of GPS-equipped smartphones, crowdsourcing has been the de facto method for real-time accident detection and road condition monitoring of many online map service providers such as Google maps [126]. Similarly, as more and more autonomous vehicles and mobile robots are deployed, crowdsourcing-based SfM may be the de facto method for creating and updating the point clouds of public outdoor and indoor environments, which can be used to guide the navigation of embodied agents. However, many issues remain to be resolved to build a practical crowdsourcing-based and SfM-powered map service for embodied navigation. First, uploading all the collected video frames is rather expensive considering the cellular data cost, so an effective strategy must be used to upload only the most crucial frames or regions of interest. Second, downloading the point cloud data in real-time for navigation challenges the bandwidth of existing network infrastructure. Possible remedies include bitrate-adaptive point cloud downloading [127,128] and point cloud compression [129]. Third, aligning reconstructed point clouds of outdoor environments with real-world dimensions (i.e., scale and orientation) can be fulfilled by using GPS coordinates [125]. However, aligning reconstructed point clouds of indoor environments should be based on other types of wireless signals such as WiFi [130–132] and Bluetooth Beacon signals [133,134], as GPS suffers significant obstructions and distortions inside buildings. Yet, wireless indoor localization itself still faces several challenges [135], including the low accuracy [136], heavy dependence on infrastructure (such as WiFi fingerprinting maps [137] and Bluetooth Beacon-enabled hardware [133,138]). We foresee the need for numerous future studies to realize the grand vision of a crowdsourcing-based and SfM-powered large-scale map service for outdoor and indoor embodied navigation.

2.2.3 *Self-localization*

In this part, we focus on the scenario where an agent tries to navigate through a known environment. Here we claim an environment to be known if it has already been explored and a corresponding map has been constructed in advance. The map may be constructed by different programs, such as the SfM program on a server and the SLAM program on other agents. Instead of mapping the environment from scratch, the agent only needs to localize itself in the readily available map. Such an offline mapping paradigm has already been adopted by many autonomous driving solutions. For example, the Baidu Apollo high-definition map [139] currently covers more than 11 million kilometers of roads across more than 360 cities and has served more than 100 million daily requests. The benefits of such a paradigm are multi-fold. First, the computation overhead on the agent is reduced as the computation-intensive mapping process is avoided. Second, different from SLAM, the map construction process in such a paradigm has a looser requirement on the throughput, so more computation-intensive construction methods such as SfM can be used to provide a more accurate map for the agent.

To achieve the goal of self-localization the agent uses the sensor to capture some data and then match the data with the offline map. Based on the involved sensor type, we divide existing studies into visual-based methods and WiFi-based methods. We do not discuss LiDAR-based methods here as they have already been discussed in a recent survey [140].

Visual methods. The conventional pipeline of visual localization contains two steps. The first step associate pixels in the query image (i.e., the image whose pose needs to be estimated) to the 3D points in the map. The correspondence can be identified through handcrafted feature descriptors [19–21] or learned feature descriptors [35, 141]. The second step involves a geometry-based pose-solver (e.g., PnP and RANSAC) to estimate the pose based on the identified correspondence.

One main issue of the above pipeline is the high latency of the correspondence matching process, which contains a few sub-steps including feature detection, feature description, and feature matching. Besides, the accuracy of the above pipeline is limited as the correspondence matching process may not always be reliable due to factors such as environmental changes. Therefore, a new approach known as scene coordinate regression (SCR) has been proposed. Instead of using feature descriptors for matching, SCR directly feeds the query image to a random forest [142–144] or a neural network [145] to obtain the per-pixel scene coordinates of the query image. Despite the higher accuracy than the conventional pipeline, SCR suffers the generalizability issue as the information about the 3D scene is encoded into the learned weights of the random forest or deep neural network (DNN). SANet [146] avoids this issue by designing a scene-agnostic network where the 3D scene is fed to the deep network together with the query image. Neither retraining nor fine-tuning is required for SANet to adapt to new scenes.

Another solution that has gained much attention recently is absolute pose regression (APR). APR aims to obtain the pose of the given query image in an end-to-end fashion and does not employ a geometry-based pose solver. Despite its conceptual simplicity, APR methods currently face two challenges. First, their accuracy is limited compared to methods incorporating a geometry-based pose solver. Second, they suffer the overfitting issue and thus require retraining or fine-tuning. The state-of-the-art method, marepo [147], still requires minutes of training to adapt to a new scene. In summary, more research is required to further improve the APR.

WiFi-based methods. In addition to being one of the most common network access technologies, WiFi has also been used by many map service providers (e.g., Google maps and AMAP [148]) to provide localization service in indoor environments, where the GPS signal is occluded and GPS-based localization is inaccurate. Compared to methods based on cameras or LiDARs, WiFi-based indoor localization has a few unique advantages. First, WiFi-based methods are free of drift accumulation as they localize WiFi devices (e.g., smartphones and WiFi-enabled robots) by using the WiFi access points (APs) as anchors. Second, WiFi-based methods are robust to complex indoor environments with illumination changes and plain visual textures. Third, WiFi-based methods have a much lower computation and energy overhead.

A common method of WiFi-based 2D localization is to utilize geometric constraints, such as the distance or the angle between the WiFi device and one or multiple APs. With such geometric constraints, the location of the WiFi device can be easily inferred through geometric algorithms such as trilateration. An analysis of the WiFi radio signal is typically required to obtain these geometric constraints. For example, the fine-grained channel state information (CSI) feature of the signal can be used to infer the distance [149]. The angle between the WiFi device and one or multiple APs can be inferred when the AP or the device has an antenna array and consecutive antennas in the array have a phase difference related to the radio signal's angle [150].

Although geometric methods are intuitive, their performance may degrade in complex indoor environments with rich multipath. Basically speaking, multipath is the phenomenon where reflectors (e.g., walls and floors) in the environment cause the radio signals to reach the receiver from multiple paths. Accurately modeling rich multipath for geometric localization is challenging, so many methods alternatively follow the idea of fingerprinting-based localization. The core idea of fingerprinting is to construct a fingerprint database that consists of fingerprints (i.e., signal features such as CSI) at different locations in the environment. The localization problem is solved by comparing the measured feature at the unknown position with fingerprints in the database to estimate the position. The process of building and updating the database is labor-intensive, and thus various methods (e.g., crowdsourcing-based methods [131, 132]) have been proposed to relieve the overhead.

Most WiFi-based methods in the literature are only capable of estimating the 2D position of the WiFi device. Some recent methods are able to estimate the position as well as the orientation of the WiFi device, which can provide more information to WiFi-enabled agents to help them better navigate through the environment. MonoLoco [151] uses a novel geometric algorithm known as multipath triangulation to simultaneously estimate its 2D position and azimuth (i.e., the angle of rotation in the 2D plane). Wi-Drone [7] achieves the 6-degree-of-freedom (6-DoF) pose estimation of the WiFi-enabled drone by designing two 6-DoF tracking algorithms and fusing the two algorithms via the factor graph.

2.2.4 *Target localization*

With the target's location perceived, the agent can take appropriate actions to navigate toward the target. Here we mainly focus on complex scenarios where the target is occluded or cannot be effectively identified by its visual appearance due to the existence of similar objects in the environment (e.g., many visually similar books in a library). Depending on the way the information about the target's position is passed to the agent, target-oriented geometric perception can be categorized into audio-based methods and methods based on RFID. As discussed later, audio-based methods are more suitable for home assistant robots while RFID-based methods are more suitable for industrial robots.

Audio-based methods. Audio-based target-oriented geometric perception typically assumes the navigation target is emitting sound and utilizes the sound signal to localize the target that may be visually unobservable due to occlusions. Audio-based perception is useful for home assistant robots as the robots can navigate to a talking human or a ringing phone. To decrease the challenge of sound source localization via deep learning, Gan et al. [152] proposed to train a sound perception module to estimate the relative position between the target and the agent. A classification model is further trained to determine whether the agent reaches the sound source. Besides, an acoustic engine is implemented to generate the sound signal in the virtual environment. Similarly, Chen et al. [153] proposed to enhance navigation by utilizing the sound signal emitted by the navigation target, but they use a more complex and more acoustically realistic simulation engine and an end-to-end navigation model that does not explicitly predict the sound source's location. Later, Chen et al. [154] proposed an omnidirectional information-gathering mechanism to collect visual-acoustic signals from many directions and achieve a large improvement in the navigation capability.

RFID-based methods. We first give a brief introduction to RFID before discussing its role in target-oriented geometric perception. A typical RFID system consists of two parts, an RFID tag and an RFID reader. An RFID tag can store some sort of identification number and the number can be retrieved by the RFID reader through radio signals. The identity of an item can be easily determined by attaching an RFID tag to the item and recording the correspondence between items and identification numbers in a database. A more detailed introduction to RFID can be found in [155]. In addition to providing the identity information, the radio signal emitted from the RFID tag also implicitly indicates the location of the tag [156–158], i.e., some characteristics (such as the phase value and the strength) of the radio signal is related to the position of the tag and can thus be used to localize the RFID-tagged item. RFID technology is very useful for target-oriented perception, especially in the context of industrial robots. First, RFID technology has already been widely deployed in various industries such as retail, manufacturing, agriculture, healthcare, and logistics. More than 44.8 billion RFID tags were shipped globally in 2023 [159], so using the existing infrastructure of RFID for target-oriented perception is feasible. In contrast, audio-based target perception may be suitable only for home robots rather than industrial robots, as items in industrial applications are generally not equipped with any sound-emitting components. Second, there exist numerous items of similar appearances in many applications, such as books in library management, clothes in a clothing store, and parcels in logistics. It is even challenging for skilled human operators to effectively distinguish these similar items by employing appearance information alone, but the problem can be easily solved through RFID technology [160]. Third, compared to cameras and LiDARs, the RFID signal can traverse everyday occlusions and enable the embodied agent to identify a partially or fully occluded navigation target more efficiently.

Many early studies use RFID alone to fulfill target-oriented geometric perception. For example, MobiTagbot [161] identifies the spatial order of multiple RFID-tagged items by roving around them and analyzing phase values collected at appropriate locations. However, as RFID signals provide limited information about the environment, using RFID alone will make it challenging for agents to achieve collision avoidance or obstacle decluttering. RF-Grasp [162] addresses this issue by fusing the RFID signal and the RGB-D image to guide the robot arm to explore the environment and grasp the target item. Compared with the baseline that does not utilize RFID, RF-Grasp improves success rate and efficiency by up to 40%–50%. Yet, RF-Grasp requires a separate RFID reader that needs to be calibrated with the robot arm. RFusion [163] works around the limitation by integrating both the RFID reader antenna and the camera onto the robot arm. Both RF-Grasp and RFusion utilize the RFID signal to localize RFID-tagged systems. However, there may exist a mixture of tagged and non-tagged items in many real-world scenarios. FuseBot [164] extends radio-based target perception to non-tagged items by excluding the positions of tagged items from consideration.

One major limitation of the aforementioned RFID-based systems is that their performances are only validated in a limited range or a constrained environment. MobiTagbot [161] requires a line-of-sight path and the orientation match between the tagged item and the RFID reader, but the requirement may not be satisfied in real-world scenarios. RFGrasp [162], RFusion [163], and FuseBot [164] assume the workspace is a small table and the robot arm is on a fixed base. A lot more research is envisioned to further improve the practicality of RFID-based target perception in complex real-world scenarios.

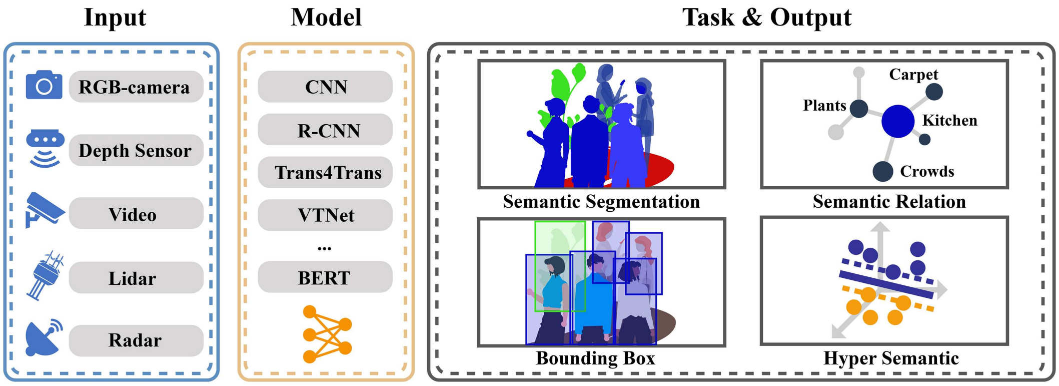

Figure 6 Semantic understanding of perception.

2.3 Semantic understanding

After constructing a geometric environment using techniques such as SLAM, semantic mapping becomes an essential process that assigns meaningful labels to different regions and objects within the environment. While the geometric map provides the physical layout, including the shapes and positions of walls, obstacles, and pathways, semantic mapping enhances this by adding contextual information. This additional layer of information allows the agent to understand what each area and object represents. By incorporating semantic labels, the agent gains a richer understanding of its surroundings, enabling it to perform more complex tasks, interact more naturally with humans, and navigate more efficiently and safely. This sector of research delves into various aspects of the semantic tasks as in Figure 6, including semantic labeling, which classifies each pixel of an image into meaningful categories; semantic relation, which identifies and understands relationships between different entities in the environment; and hyper-semantic enabled by multi-modal semantic integration, which combines data from various sensors to create a comprehensive and cohesive understanding of the environment. Together, these advancements enable more sophisticated and intuitive interactions, making semantic mapping an indispensable component of modern navigation systems.

2.3.1 *Semantic labeling*

Semantic segmentation plays a crucial role in embodied navigation, where an agent must interact intelligently with its environment. By assigning a class label to each pixel in an image, semantic segmentation allows the agent to understand the layout of the scene, identify objects, and distinguish between navigable and non-navigable areas. This granular understanding is essential for tasks like path planning, obstacle avoidance, and object interaction, enabling the agent to navigate efficiently and safely in complex, dynamic environments. Initially, semantic segmentation predominantly utilized convolutional neural network (CNN)-based methods [3, 165–167] for object detection and scene understanding. However, with the advent and rapid advancement of transformer architectures, the current mainstream approaches for semantic segmentation have increasingly shifted towards leveraging transformers. These architectures have demonstrated superior performance in capturing long-range dependencies and contextual information, which are crucial for accurate semantic segmentation tasks. In the following section, we will delve into the semantic segmentation techniques that are based on transformer architectures, highlighting their methodologies and the improvements they bring to the field.

Transformer architectures have revolutionized semantic segmentation by adapting to image inputs, as demonstrated by Dosovitskiy et al. [168]. This adaptability has enabled the use of transformers for semantic segmentation, allowing for the effective processing of image patches to capture long-range dependencies and contextual information, which are crucial for accurate segmentation. Later, Carion et al. [169] demonstrated that transformer-based methods outperform traditional CNN and R-CNN approaches. Transformers excel in modeling global relationships within the image, leading to superior performance in object detection and segmentation tasks. This improvement is attributed to the self-attention mechanism in transformers, which captures intricate details and relationships between different parts of the image more effectively than CNN-based methods. Furthermore, Zhang et al. [170] introduced the Trans4Trans model, which highlights the capability of transformer-based methods to achieve better performance in detecting challenging transparent objects. This is particularly beneficial for applications such as navigation assistance for visually impaired individuals, where identifying transparent obstacles like glass doors is critical for safety. The Trans4Trans model's dual-head design effectively segments both general and transparent objects, showcasing the versatility and robustness of transformer architectures in real-world scenarios.

2.3.2 *Semantic relation*

Semantic segmentation provides a foundational layer for autonomous agents to comprehend and navigate their environments by partitioning images into semantically meaningful regions, enabling agents to distinguish between various objects and surfaces. However, beyond this initial step, there exist advanced methodologies that leverage semantic segmentation data to further enhance the understanding of relationships between objects, thereby optimizing navigation efficiency. The ability to understand object relationships provides several merits, such as predicting the presence of certain objects based on the detection of related objects, making navigation more intuitive and effective, and improving the overall robustness of the navigation system.

Wu et al. [171] introduced a method that combines Bayesian relational memory with semantic mapping to improve navigation tasks. The system maintains a probabilistic map that captures relationships between objects and their locations. By integrating semantic information within a memory framework, the agent can make more accurate predictions about the presence and positions of objects. This probabilistic approach facilitates better path planning and decision-making, leading to more efficient navigation. The relationships between objects, captured probabilistically, help the agent to navigate more effectively by anticipating the locations of related objects.

In the study by Mousavian et al. [172], high-level visual representations derived from semantic segmentation and detection masks are used to train DNNs, capturing spatial layouts and contextual cues. These networks achieve robust generalization and enable agents to navigate towards specified target objects efficiently by understanding the relationships between segmented objects. Similarly, Yang et al. [173] incorporated semantic priors using graph convolutional networks (GCNs) to encode knowledge about object relationships and typical placements within environments. This predictive capability allows the agent to make informed decisions about where to search for objects based on observed scene context, even in previously unseen environments. Furthermore, in the study by Du et al. [174], transformer architectures are leveraged to process image inputs, capturing both local and global contextual information. The self-attention mechanism in transformers allows the agent to identify detailed object features and their relationships within the scene, enabling high accuracy in navigation even in complex environments. By modeling the relationships between objects through these advanced techniques, these studies collectively demonstrate how semantic information can be used to build a comprehensive understanding of the environment, facilitating more effective and efficient navigation.

2.3.3 *Hyper-semantic*

In recent years, advancements in multi-modal sensor technologies, including visual, auditory, and linguistic inputs, have significantly enhanced our ability to collect and integrate diverse types of data, thereby enabling more accurate and comprehensive semantic features of the environments other than object categories/relations. We refer to this as hyper-semantic. This fusion of data from multiple sensory modalities represented in latent space makes it possible to create a more holistic understanding of the environment, which is critical for the development of robust and adaptive intelligent agents.

The integration of vision and language represents a pivotal direction of hyper-semantic application in embodied navigation, known as vision-and-language navigation (VLN), synthesizes advancements in computer vision and natural language processing, aiming at equipping agents with the capability to comprehend and execute complex natural language instructions within visually rich environments. Shah et al. [2] introduced LM-Nav, a system that leverages large pre-trained models for navigation, vision, and language without requiring fine-tuning in the target environment. LM-Nav utilizes models such as GPT-3 for language parsing, CLIP for vision-language association, and ViNG for visual navigation, demonstrating robust performance in complex, real-world environments. Similarly, Hao et al. [175] proposed a pre-training and fine-tuning paradigm for VLN, using large datasets of image-text-action triplets to improve generalization across various environments. In a related effort, the episodic transformer introduced by Pashevich et al. [176] employed transformer models to handle long-term dependencies in

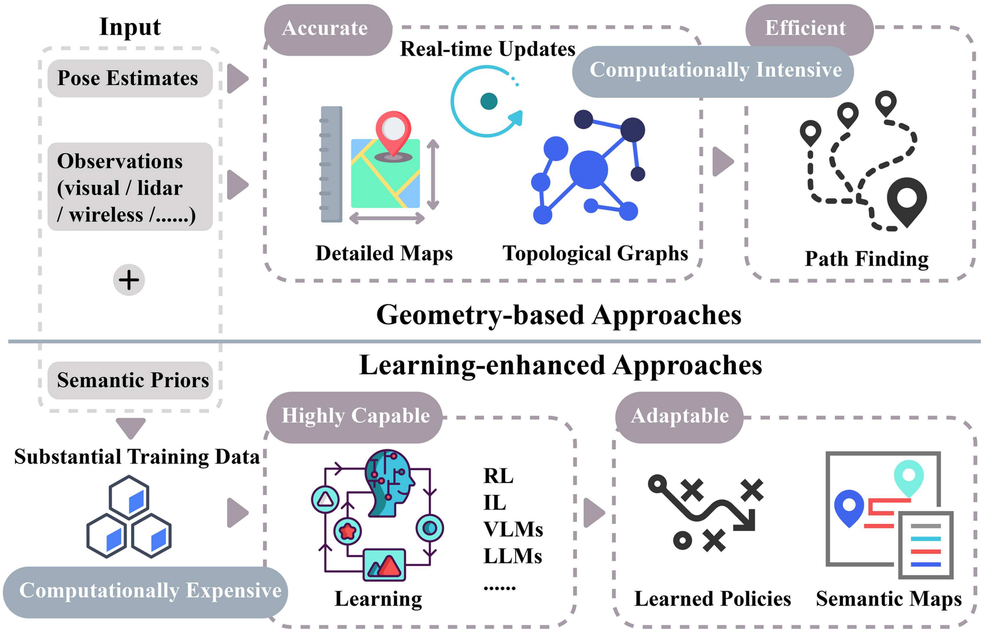

Figure 7 Geometry-based and learning-enhanced approaches.

vision-and-language tasks, further advancing the capabilities of VLN systems. Additionally, VLN research has explored the integration of embodied AI for task completion. For instance, the embodied BERT model introduced by Suglia et al. [177] is designed to attend to high-dimensional, multi-modal inputs for language-conditioned task completion, demonstrating significant improvements in performance on tasks like those in the ALFRED benchmark.

Leveraging audio information for hyper-semantic has shown significant improvements in navigation performance. For instance, Gan et al. [152] proposed a model where an agent integrates audio and visual data to navigate toward a sound source, employing a visual perception mapper and a sound perception module to construct spatial memory and infer sound locations. Similarly, Chen et al. [178] introduced a method where the agent uses semantically meaningful audio cues, such as a door creaking, to guide navigation, combining these cues with visual observations to maintain direction even after the sound stops. Furthermore, the AVLEN framework by Paul et al. [179] focused on using audio, visual, and language inputs in a multimodal approach, allowing the agent to localize audio events and seek assistance from a human oracle through natural language when necessary. These methods collectively demonstrate that incorporating audio data into navigation systems significantly enhances the agent's ability to perform complex tasks in dynamic environments. Additionally, the non-local fusion network (NLFNet) proposed by Yan et al. [180] demonstrated the effectiveness of selectively fusing multimodal input information, including RGB, depth, polarization, and thermal images into hyper-semantic information. This approach improved the segmentation accuracy by leveraging the complementary information provided by different optical sensors, thereby addressing the problem of object recognition in various challenging real-world scenes. This work highlights the importance of integrating different optical information to complement visual data, further enhancing the robustness of semantic identification.

3 Navigation

Embodied navigation is the critical component of embodied AI, and involves guiding agents through physical environments to achieve specific goals. This section delves into various methodologies employed in embodied navigation, categorizing them into geometry-based and learning-based approaches as in Figure 7. Each approach brings unique advantages and challenges, reflecting the diverse strategies that researchers employ to enhance navigation capabilities in embodied agents. We have provided a brief comparison between the two types of approaches in Table 1.

Table 1 Comparison of geometry-based and learning-enhanced approaches in embodied navigation.

Approach	Geometry-based	Learning-enhanced
Key techniques	Map-based, graph-based	Reinforcement learning (RL), imitation learning (IL), vision-language models (VLMs), large language models (LLMs)
Environment understanding	Detailed metric maps, topological graphs, limited semantic maps	Learned policies, semantic maps enriched with pre-trained models
Adaptability	Moderate (predefined structures, some adaptability through dynamic updates)	High (learning from interactions)
Computational requirements	High for real-time updates of maps and graphs	High for model training and inference
Strengths	Accurate spatial understanding, efficient pathfinding	Flexible, adaptable, capable of handling complex tasks, utilizes rich semantic priors
Weaknesses	Limited semantic understanding, computationally intensive for real-time updates	Requires significant training data, computationally expensive

3.1 Geometry-based approaches

Geometry-based navigation methods rely on constructing and updating spatial representations of the environment in real-time, without the use of reinforcement learning (RL) techniques. These methods focus on creating and refining maps and graphs as the agent explores the environment, leveraging the geometrical and topological structure to navigate efficiently. The core idea is to enable agents to build a comprehensive understanding of their surroundings dynamically, facilitating effective navigation through structured spatial representations.

3.1.1 *Map-based navigation*

Map-based navigation involves the creation and continuous updating of detailed maps that provide spatial information about the environment. These maps, which can be occupancy grids enriched with semantic information, are constructed dynamically as the agent explores. This process helps the agent understand the spatial layout and locate objects within the environment. Occupancy grid maps partition the environment into discrete cells, each indicating whether a location is occupied, free, or unknown. These maps can also be enhanced with semantic categories and probabilities, offering a nuanced and comprehensive representation of the environment.

The evolution of map-based navigation techniques can be traced through a series of seminal studies that laid the foundation for this approach, followed by more recent advancements that have built upon these early efforts.

The concept of occupancy grid mapping was first introduced by Elfes [1], providing a robust method for handling sensor data and map construction. This approach used a probabilistic tessellated representation of spatial information, which allowed the robot to integrate data from multiple sensors and different viewpoints to incrementally update a coherent map.

Building on the foundations laid by occupancy grids, researchers explored efficient pathfinding algorithms to optimize navigation. One of the earliest and most influential algorithms is the A* algorithm [181], which combines features of Dijkstra's algorithm and best-first search to find the shortest path efficiently. A* uses heuristics to guide its search, making it faster and more efficient than Dijkstra's algorithm alone. Recognizing the need for adaptability in dynamic environments, Stentz [182] introduced the D* algorithm, an extension of A* specifically designed to enable real-time updates of maps and paths. This advancement was crucial for dynamic path planning, allowing robots to respond to changes in the environment as they occurred. In parallel, the SLAM framework introduced by Leonard and Durrant-Whyte [183] addressed the challenge of enabling robots to build maps of unknown environments while determining their own location. This approach allowed for the concurrent construction of a map and localization of the robot, a critical capability for autonomous navigation in uncharted territories.

As the field progressed, the integration of probabilistic methods became increasingly important. Thrun, Burgard, and Fox's "probabilistic robotics" [184] systematically presented probabilistic methods for robot localization and mapping, including detailed discussions on occupancy grids. This comprehensive text became essential for understanding and implementing map-based navigation techniques, emphasizing the importance of probabilistic approaches in handling uncertainties and integrating sensor data.

Recent advancements have significantly refined map-based navigation by integrating modern computational methods and richer datasets. For instance, Luo et al. [185] proposed the Stubborn method, which introduced a semantic-agnostic exploration strategy and multi-scale collision maps, addressing key challenges in exploration and object recognition within map-based frameworks. This approach demonstrated that even without semantic understanding, robust navigation could be achieved through careful exploration and collision avoidance techniques.

Similarly, Ramakrishnan et al. [186] developed the PONI method, a map-based approach leveraging potential functions to decompose the navigation problem into manageable components, reducing computational requirements while maintaining effective navigation capabilities. This method showed how separating perception and movement can lead to more efficient navigation strategies.

Continuing this trend, Georgakis et al. [187] employed an innovative active learning framework in their map-based method for creating and utilizing semantic maps. By actively selecting training samples to maximize information gain, they significantly improved navigation efficiency. Their approach leveraged uncertainty estimation and an upper confidence bound strategy to balance exploration and exploitation, demonstrating superior performance in unknown environments and highlighting the benefits of combining traditional mapping with active learning to enhance goal-directed navigation.

Also, the map-based approach by Zhu et al. [188] used semantic cues to predict distances to target objects, selecting optimal intermediate goals and improving navigation paths. This method showcased the potential of integrating semantic understanding with distance prediction to navigate efficiently in previously unseen environments.

These advancements illustrate the continuous evolution of map-based navigation techniques. By building on foundational principles and integrating modern computational methods, researchers have significantly enhanced the robustness, efficiency, and capabilities of autonomous navigation systems. This progression from early occupancy grid maps to sophisticated, dynamic, and semantically enriched navigation strategies demonstrates the field's ongoing innovation and adaptation to complex environments.

3.1.2 *Graph-based navigation*

Graph-based navigation methods represent the environment as a network of nodes and edges, where nodes correspond to key locations or landmarks and edges denote navigable paths between them. This topological representation allows agents to navigate by following paths through the graph, making decisions at each node based on the available edges. Unlike map-based approaches that rely heavily on metric maps and detailed spatial representations, graph-based methods focus on the connectivity and relationships between different locations. This approach allows for more abstract and flexible navigation strategies, especially in large and complex environments.

One of the early seminal studies in graph-based navigation is Kuipers's "Modeling spatial knowledge" (1978) [189], where he introduced the TOUR model. This model constructs cognitive maps that humans use to navigate large-scale spaces through a network of interconnected nodes and edges. Kuipers' work laid the foundation for using graph structures in navigation, systematically formalizing the idea of representing environments with nodes as key locations and edges as paths. This approach emphasized the importance of partial knowledge states and gradual learning, significantly influencing subsequent research in topological mapping and graph-based navigation.

Building on these foundational ideas, Thrun and Bücken [190] advanced graph-based navigation by integrating grid-based and topological maps. They proposed a hybrid approach that constructs grid-based maps using neural networks and Bayesian integration, then generates topological maps by partitioning the grid into coherent regions connected by critical lines. This method effectively creates a graph structure overlaid on a metric map, leveraging the precision of grid-based representations and the efficiency of topological graphs. By combining these paradigms, their approach enhances the robot's ability to plan and navigate autonomously, enabling real-time navigation in complex environments with improved accuracy and efficiency.

To further enhance the capabilities of graph-based navigation, recent advancements have introduced GCNs. GCNs extend the concept of CNNs to graph-structured data, allowing for efficient processing and analysis of non-Euclidean data. Introduced by Kipf and Welling [191], GCNs have been instrumental in various applications, including navigation tasks. By leveraging the structural properties of graphs, GCNs enable the extraction and utilization of rich relational information, facilitating more sophisticated and effective navigation strategies.

A notable application of GCNs in graph-based navigation is by Kiran et al. [192], who use spatial relation graphs (SRG) and GCNs for object-goal navigation. Their framework learns an SRG from the robot's trajectories, modeling the proximity between regions and object occurrences. Bayesian inference and GCN-based embeddings are then used to rank and select regions for exploration. Tested in the Matterport3D benchmark within the AI Habitat environment, this method significantly improves navigation accuracy and efficiency over state-of-the-art baselines.

Building on this, Liu et al. [193] proposed the ReVoLT framework, combining relational reasoning with Voronoi local graph planning. This hierarchical approach uses GCNs and other methods to infer semantic sub-goals and plan paths. Experiments in the AI Habitat environment show that ReVoLT significantly outperforms existing methods.

The discussed studies highlight the evolution and refinement of graph-based navigation methods. These approaches, which model the environment as interconnected nodes and edges, provide a flexible framework for navigation, particularly in extensive and intricate environments. The advent of techniques like GCNs has significantly improved the ability of agents to process and utilize relational information within these graphs, thereby enhancing their navigation performance. These advancements underscore the robustness and adaptability of graph-based navigation in the realm of embodied AI, paving the way for more sophisticated and capable navigation systems.

3.2 Learning-enhanced approaches

Learning-enhanced navigation approaches harness the power of machine learning to enable agents to learn optimal navigation strategies through interaction with their environment. Unlike geometry-based approaches, which rely heavily on constructing and updating spatial representations, learning-enhanced methods focus on the agent's ability to adapt and improve its navigation policy through continuous learning and experience. This subsection explores various learning-enhanced navigation strategies, emphasizing their distinctive reliance on learning for decision-making rather than solely on predefined geometric structures.

3.2.1 *RL & imitation learning (IL)*

RL and IL are powerful techniques that enable agents to develop sophisticated navigation policies through interactions within their environment. RL methods involve training agents to maximize cumulative rewards through trial and error, while IL involves mimicking expert demonstrations to achieve desired outcomes. Unlike the geometry-based methods discussed in Subsection 3.1, which focus on creating detailed spatial representations, RL and IL emphasize learning from interactions and experiences to improve navigation strategies. We delve into the application of RL and IL in navigation tasks, highlighting key advancements and methodologies that have enhanced the agent's ability to navigate complex environments.

Moving forward from the foundational studies, the study by Chaplot et al. [167] introduced the SemExp method. This method builds an episodic semantic map using existing object detection and semantic segmentation models, enabling the agent to navigate efficiently based on the goal object category by leveraging semantic priors and long-term episodic memory. While earlier studies had incorporated semantic understanding in navigation tasks, this pioneering approach significantly advanced the field by demonstrating superior performance compared to a wide range of baselines.

Building on these early foundations, Ye et al. [194] explored the integration of auxiliary tasks and exploration rewards to enhance object-goal navigation performance. By designing a series of auxiliary tasks, the agent could better understand and explore its environment, thereby improving its navigation efficiency. This study demonstrated the potential of multi-task learning and exploration rewards in optimizing navigation strategies, marking another step forward in the evolution of RL and IL applications.

As the field continued to evolve, addressing the complexity of exploration tasks required innovative solutions. To tackle the challenges of engineering effective reward functions and scalability, Ramrakhya et al. [195] developed the Habitat-Web framework. This framework utilized a web-based application to collect large-scale human demonstration data via crowdsourcing. The use of real human demonstrations allowed the agent to learn complex navigation and manipulation tasks through IL, enhancing the robustness and generality of the learned policies. This approach highlighted the importance of human data in advancing navigation techniques.

In parallel, addressing the issue of transferability and zero-shot learning in visual navigation, Al-Halah et al. [196] proposed a plug-and-play modular transfer learning framework. This framework allowed agents to leverage pre-trained models for various navigation tasks without requiring additional training data. This method significantly improved training efficiency and task generalization, enabling agents to perform zero-shot learning on new navigation tasks, thus bridging the gap between pre-trained models and new environments.

Combining RL with IL to overcome the limitations of behavior cloning (BC) marked another significant advancement. Ramrakhya et al. [197] developed the PIRLNav approach, which involved pretraining policies using IL and fine-tuning them with RL. This combination addressed the poor generalization of BC strategies and the high cost of collecting demonstration data, resulting in a more adaptable and efficient navigation policy. This hybrid approach demonstrated the synergy between RL and IL, further pushing the boundaries of navigation capabilities.

Further advancements in 3D-aware navigation emphasized the importance of learning from fine-grained spatial information in 3D environments. Zhang et al. [198] proposed an approach that integrated simultaneous exploration and identification, allowing agents to construct detailed 3D environment models and significantly enhance object-goal navigation performance. This innovation underscored the necessity of detailed spatial understanding in achieving high navigation efficiency.

To address the challenge of limited interaction with experts, Singh et al. [199] introduced the Ask4Help framework. This framework enabled agents to request help from experts when needed, allowing efficient training without altering the original agent parameters. It achieved a balance between task performance and expert assistance requests, reducing the overall cost of expert involvement. This method illustrated the practical approach of leveraging expert knowledge to enhance training processes.

The aforementioned studies illustrate the significant advancements in applying RL and IL to navigation tasks. Through the continuous improvement of learning algorithms and optimization strategies, RL and IL have substantially enhanced the capability of agents to navigate complex environments. These methods have proven essential in enabling agents to learn from interactions and experiences, thereby developing robust and efficient navigation policies that can adapt to various challenges in embodied navigation.

3.2.2 *Vision-language models (VLMs) & large language models (LLMs)*

VLMs and LLMs integrate visual processing and natural language understanding to enhance navigation capabilities. These models use large-scale pre-trained networks to interpret and generate both visual information and natural language descriptions, enabling agents to follow complex instructions and make informed decisions. Unlike geometry-based approaches that construct spatial representations and RL/IL methods that learn from interactions, VLMs and LLMs combine semantic understanding with visual cues.

VLMs such as CLIP [200], ALIGN [201], and FLAVA [202] have enhanced visual semantic understanding by integrating visual and language features. Similarly, LLMs like GPT-3 [203], T5 [204], and BERT [205] have demonstrated powerful capabilities in understanding and generating natural language. The introduction of GPT-4 has further pushed the boundaries of what LLMs can achieve, with even more sophisticated language processing abilities. These breakthroughs collectively enable the effective application of VLMs and LLMs in complex navigation tasks. We explore how VLMs and LLMs, both individually and together, facilitate navigation in complex environments.

Recent advancements in VLMs have demonstrated significant potential in zero-shot navigation. Majumdar et al. [206] introduced an approach using CLIP to create a visiolinguistic embedding space, enabling agents to navigate in an open-world setting. This method addresses the closed-world assumption in traditional ObjectNav by allowing agents to understand and navigate toward objects described in arbitrary natural language terms, such as "flat-head screwdriver" or "bathroom sink". Building on the concept of zero-shot capabilities, another method by Zhou et al. [207] further enhances navigation by integrating soft commonsense constraints. This approach leverages multimodal goal embeddings created using the CLIP model, improving the agent's performance in unfamiliar settings and enabling robust and flexible navigation in open-world environments.

LLMs have also made significant strides in navigation tasks through their advanced language understanding and reasoning capabilities. The framework proposed by Zhou et al. [208] utilized the explicit reasoning capabilities of GPT-4 to handle complex VLN tasks. This model excels in intricate planning, common-sense reasoning, and handling unexpected events, thereby enhancing the agent's ability to navigate through detailed instructions and dynamic environments. Complementing this, another framework

Table 2 Comparison of success rates and SPL for object-goal and vision-language navigation methods.

Method	Dataset	Success rate (%)	SPL	Dataset	Success rate (%)	SPL
Stubborn [185]	MP3D (val)	23.7	0.098	MP3D (test-standard)	23.7	0.098
PONI [186]	MP3D (val)	31.8	12.1	MP3D (test-standard)	20.01	8.82
L2M-Active-Upper [187]	MP3D (val)	34.3	13.3	–	–	–
DP (distance prediction) [188]	MP3D (val)	43.8	23.2	–	–	–
SRG-GCN [192]	MP3D (val)	75.1	53.0	–	–	–
ReVoLT [193]	MP3D (val)	66.7	26.5	–	–	–
SemExp [167]	MP3D (val)	36.0	14.4	MP3D (test-challenge)	25.3	10.2
AuxTask [194]	MP3D (val)	34.4	9.58	MP3D (test-standard)	24.5	8.1
Habitat-Web [195]	MP3D (val)	35.4	10.2	MP3D (test-standard)	27.8	9.9
Plug&Play [196]	MP3D (test)	14.6	10.8	HM3D (test)	9.6	6.3
PIRLNav [197]	HM3D (val)	61.9	27.9	HM3D (test-standard)	65.0	33.0
3D-Aware [198]	Gibson (val)	74.5	42.1	MP3D (val)	34.0	14.6
Ask4Help [199]	RoboTHOR (val)	81.6	24.3	–	–	–
ZSON [206]	Gibson (val)	31.3	12.0	HM3D (val)	25.5	12.6
ESC [207]	MP3D (val)	28.7	14.2	HM3D (val)	39.2	22.3
NavGPT [208]	R2R (val unseen)	34.0	29.0	–	–	–
VELMA [209]	Touchdown (val)	23.4	–	Map2seq (val)	41.3	–
VLMaps [210]	MP3D (val)	59.0	34.0	–	–	–
LM-Nav [2]	Real-world	80.0	–	–	–	–

by Schumann et al. [209] explores the embodiment of LLM agents for vision and language navigation in street views. By emphasizing verbalization, this method allows the agent to effectively interpret and execute navigation instructions in street-level environments, demonstrating practical applications of LLMs in real-world navigation scenarios.

Integrating VLMs and LLMs offers a synergistic approach that leverages the strengths of both models. Huang et al. [210] introduced a method that creates visual language maps (VLMaps) for robot navigation by combining VLMs and LLMs. The VLM generates visual-language embeddings from the robot's video feed, while the LLM processes instructions to ground these goals in the visual context, enhancing understanding and navigation precision. This method allows robots to follow detailed spatial instructions, enabling precise navigation in complex environments. Shah et al. [2] developed the LM-Nav framework, which employs pre-trained models of language, vision, and action for robust robotic navigation. This framework uses CLIP as the VLM to associate images with textual landmarks and GPT-3 as the LLM to parse and translate instructions. By correlating image-language data and executing actions based on this combined understanding, LM-Nav bridges the gap between visual comprehension and language reasoning. This integrated approach enhances the robot's ability to handle complex, real-world navigation tasks, demonstrating the powerful synergy between VLMs and LLMs.

The integration of VLMs and LLMs significantly enhances the capabilities of navigation systems by combining visual semantics with advanced language processing. VLMs excel in creating rich visual-language embeddings that allow for zero-shot navigation in open-world settings, while LLMs provide powerful reasoning and language understanding for interpreting complex instructions. Together, these models enable robust and flexible navigation strategies, as demonstrated by approaches like VLMaps and the LM-Nav framework. These methods highlight the synergistic potential of VLMs and LLMs to bridge the gap between visual and linguistic understanding, paving the way for more adaptable and intelligent navigation systems capable of operating in diverse and dynamic environments. In Table 2 [2,167,185–188,192–199,206–210], we have listed the success rates and success weighted by path length (SPL) of mainstream object-goal and vision-language navigation methods for comparison. From the results, we observe that methods incorporating semantic priors, exploration strategies, or human demonstrations generally achieve better performance, highlighting the significance of leveraging high-level scene understanding and data efficiency in embodied navigation tasks.

4 Efficiency optimization for embodied navigation

Efficiency is important in the design and operation of embodied navigation systems. These systems must navigate through complex environments while responding to changes in real time. To achieve this, several key factors must be optimized, including latency, energy efficiency, and robustness. Each of these factors plays a vital role in ensuring that the system can perform its tasks safely and effectively. This section will discuss challenges and strategies for optimizing these aspects.

4.1 Latency optimization

To ensure safety, embodied navigation agents often have a strict budget for end-to-end latency. For drones, autonomous cars, and robots, high latency can result in missed opportunities to avoid obstacles or take optimal paths, which is crucial for safe and efficient navigation. Besides, Efficient navigation involves continuous path planning and re-planning. Low latency ensures that the agent can quickly adjust its path based on new information, optimizing travel time and energy consumption. Therefore, optimizing end-to-end latency is crucial for embodied navigation tasks.

However, ensuring real-time reaction for embodied navigation agents involves several challenges. First, embodied navigation agents often involve multiple tasks such as perception, control, and decision-making, many of which typically rely on computationally intensive algorithms such as DNNs, which can cause high latency. Achieving high computational efficiency with limited resources is a challenge in embodied navigation [211]. Second, the environment of an embodied agent is dynamic, with factors like temperature variations, hardware aging, and background processes. This requires edge devices to switch applications or models to handle new tasks in changing environments [212]. Thus, maintaining stable computation latency in dynamic environments is another significant challenge [213, 214]. Third, network and communication challenges play a crucial role in ensuring real-time performance. Systems relying on remote processing or cloud services can face significant network latency, requiring fast and reliable communication links. There have been multiple techniques proposed to address the challenges.

4.1.1 *Adaptive computation*

Adaptive computation is to dynamically adjust computational resource allocation based on the real-time demands of the navigation task as shown in Figure 8. This ensures that the latency budget is met even in dynamic environments, as highlighted by various studies [213–215]. MCDNN [212] achieves this through model scheduling, employing an optimizing compiler and runtime scheduler to facilitate the execution of multiple DNN models across mobile and cloud devices, thereby optimizing resource utilization and enhancing overall system performance. NestDNN [213] introduces an on-device adaptive deep learning framework that operates independently of cloud connectivity, accounting for runtime resource variability to enable resource-aware DNN models tailored for mobile vision systems. LegoDNN [215] introduces a novel block-grained DNN model scaling scheme. This scheme is tailored for running multiple DNN workloads on mobile devices, offering a modular approach that allows for efficient scaling and management of DNN models. AdaptiveNet [214] proposes an innovative model elastification method designed to adapt DNN architectures post-deployment. Compared to former studies, AdaptiveNet could adapt DNN architecture after deployment, where the model quality could be precisely measured. These studies provide robust solutions for ensuring latency budget in dynamic and resource-constrained environments.

4.1.2 *Parallel processing*

Leveraging multiple processors and specialized hardware accelerators to distribute computational tasks can also enable faster processing and response times [216, 217]. With various processing units like CPUs, GPUs, and NPUs, mobile devices such as robots and intelligent vehicles can decrease inference latency through parallel execution on different processors [218]. Efficient parallel methods have been explored to optimize computation distribution [219], balance the load [220], and minimize communication costs across processors [221]. BioDrone [8] ensures real-time reaction by leveraging parallel computation through FPGA-based implementations of the chiasm-inspired event filtering (CEF) and lateral geniculate nucleus (LGN)-inspired event matching (LEM) algorithms. These algorithms process event data in parallel, significantly reducing latency. DNNFusion [222] introduces advanced operator fusion techniques and efficient code generation, which reduces memory pressure and improves execution efficiency, leading to

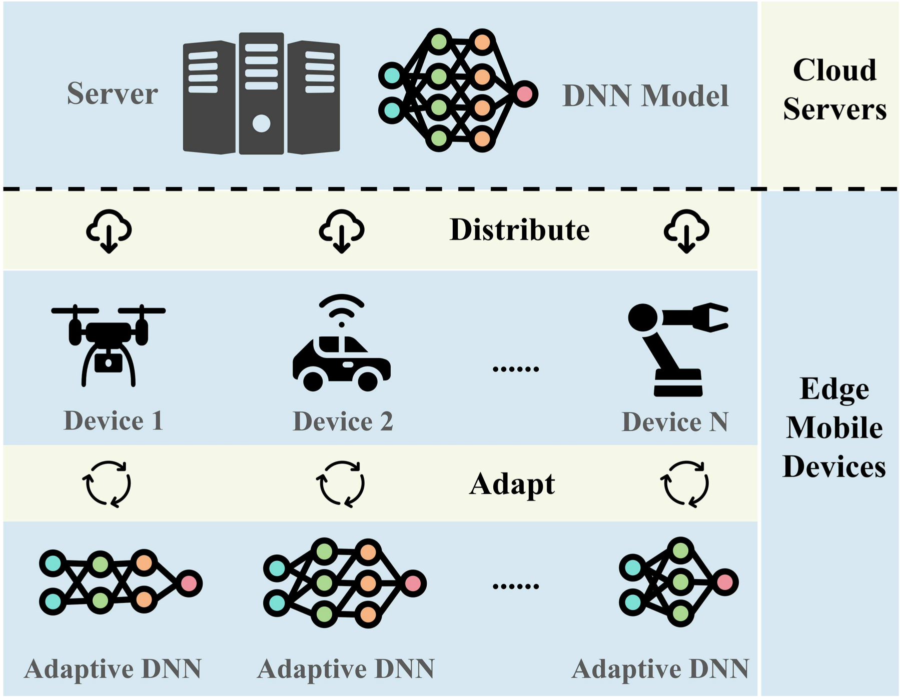

Figure 8 Adaptive computing.

high parallelism and low latency of DNN models. NN-Stretch [219] enhances computation parallelism in deep learning models by converting them into multibranch structures. It achieves this by transforming traditional single-branch models into shorter and wider models with multiple branches, optimizing them for execution on heterogeneous multi-processors.

4.1.3 *Communication optimization*

For embodied navigation systems that rely on remote services or communication with other agents, efficient communication is vital to maintaining low latency. First, edge computing can alleviate latency issues associated with cloud dependence. By processing data closer to the source, edge computing reduces the time required for data to travel back and forth between the device and the cloud. EdgeSLAM [33] uses a novel computation offloading strategy and an adaptive task scheduling algorithm to achieve real-time SLAM-based navigation. The system also considers dynamic network conditions and adjusts system parameters accordingly. SwarmMap [6] presents a framework that scales up collaborative SLAM-based agents in edge offloading settings. It can achieve real-time collaboration between numerous agents by implementing a map information tracker, an SLAM-specific task scheduler, and a map backbone profiling module. Second, optimizing data transmission volume is also critical. Techniques like data compression [223] and selective data transmission [224] can significantly decrease the amount of data that needs to be sent over the network, thereby reducing latency. CoEdge [225] introduces a collaborative edge system tailored for distributed real-time deep learning tasks, enhancing data transmission efficiency by batching sensor inputs and aggregating inference results. VILAM [226] optimizes data transmission by extracting and transmitting a compact representation of the static environment, extracting low-frequency measurements, and utilizing parallel processing techniques. This reduces data volume and ensures efficient real-time communication and localization.

4.2 Energy efficiency optimization

The compelling need for better power and energy sources is one of the most major challenges to the advancement of autonomous systems. Particularly for drones and mobile robots, their battery storage capacity is limited due to size constraints. Additionally, the navigation tasks often need to be performed in harsh environments, which increases operational complexity. Long-duration missions place higher demands on energy efficiency, which is particularly crucial for extended operations. Therefore, improving the energy efficiency of path planning is crucial for the performance of embodied navigation tasks.

However, improving energy efficiency in path planning for robots involves several significant challenges. We summarize three main challenges.

- Complex environments and obstacles. Path planning algorithms must constantly adapt to changes, which can be computationally intensive and energy-consuming.
- Algorithmic efficiency. Many traditional path planning algorithms, like those based on grid maps or sampling-based methods, may not be inherently energy-efficient. They often focus on finding the shortest or safest path without considering the energy cost explicitly.
- Real-time processing. Achieving real-time processing while optimizing for energy efficiency is difficult, especially for autonomous mobile robots operating in dynamic environments. This requires advanced computing techniques and often specialized hardware to ensure that energy-efficient paths can be computed and adjusted on the fly.

Addressing these challenges involves a multidisciplinary approach, integrating advancements in algorithm design, robotics hardware, and energy management systems. Many studies have proposed corresponding solutions to improve energy efficiency in path planning for embodied navigation. These include Bayesian algorithms, genetic algorithms, and RL algorithms for single-agent systems, as well as multi-stage heuristic algorithms for multi-agent path coverage problems. Specifically, the path planning problem of individual intelligent agents has rich applications in scenarios such as autonomous electric vehicles, UAVs, and mobile robots. Ref. [227] integrated a hybrid evolutionary algorithm and Q-learning, taking into account both drone speed and distance to obstacles. To overcome the limitations of traditional optimization methods, this approach combines genetic algorithms with Q-learning. By considering drone speed and distance to obstacles, the method optimizes path planning decisions based on real-time information. Ref. [228] proposed a multi-agent path planning scheme based on deep RL. The objective is to minimize the total energy consumption of drones while completing unloading tasks. Building on segmented aerial images of the environment as introduced in [229], it demonstrates the establishment of a rough energy map from segmentation. In the offline phase, this map is utilized to construct a covariance function for environmental representation based on Gaussian processes (GP). In the online phase, energy measurements collected during navigation are employed to estimate the energy distribution of the entire environment using GP regression, thereby reducing energy consumption during drone navigation. Additionally, the collaborative task processing of multiple intelligent agents is also a very common scenario, such as drone network communication. The energy optimization work in this area is also very important. The most important path planning problem for drone swarms is to consider the use of multiple drones to cover any area containing obstacles, known as the coverage path planning (CPP) problem. The objective of the CPP problem is to find paths for each drone to cover the entire area. Ref. [230] employed Bayesian methods to model the energy consumption of road segments for efficient navigation. An online learning framework was developed to learn model parameters, investigating various exploration strategies. This framework was then extended to multi-agent environments, allowing multiple vehicles to dynamically adjust navigation strategies and learn energy model parameters based on environmental changes. Real-world experiments on the Luxembourg SUMO traffic dataset demonstrate the energy optimization performance of the approach. Ref. [231] tackled the CPP problem in two steps: first, distributing the given area's workload evenly among individual drones, and second, solving the minimum energy path planning (MEPP) problem for each drone. It introduces a well-known Lin-Kernighan heuristic method to effectively address the issue.

4.3 Robustness improvement

Robustness in embodied navigation specifically refers to the ability of an embodied AI agent (such as a robot) to consistently and effectively navigate through its environment despite facing a range of challenges. These challenges can include changes in the task environment, and even sensor noise. Robust embodied navigation ensures that the agent can adapt to these variations and maintain its navigation performance, achieving its goals without failure or significant performance degradation. Therefore, robustness is a crucial characteristic of an embodied navigation system, requiring computationally efficient and intensive strategies to ensure generalization. Next, we will introduce methods for improving robustness from two perspectives: adaptation to environmental changes and enhancement of sensor reliability.

4.3.1 *Adaptation to environmental changes*

An environment change refers to the fact that deployment environments possess different appearance statistics (such as weather changes) and motion dynamics (such as varying ground friction coefficients) compared to the environments used for training those agents [232]. Chattopadhyay et al. [232] proposed a framework, RobustNav, as a first step to assess the robustness of embodied agents against various visual and dynamics corruptions. Their findings reveal that standard agents often underperform under such corruptions, highlighting the need for robustness techniques such as data augmentation and self-supervised adaptation [233]. RobustNav serves as a testbed for enhancing navigation performance in diverse perceptual and actuation conditions, and emphasizes that zero-shot adaptation approaches can be more efficient and scalable than adapting to every possible target scenario due to the enormous variety of unseen environments and situations.

Peng et al. [234] proposed RDMAE-Nav, a navigation framework for enhancing the robustness of embodied agents during PointGoal navigation in the presence of visual corruptions. RDMAE-Nav integrates a visual module employing regularized denoising masked autoencoders (RDMAE) for robust representation learning, along with a policy module. It achieves competitive performance compared to state-of-the-art methods, showcasing significant enhancements under diverse visual corruptions.

Recently, a novel adaptation scheme, test-time adaptation, has been applied to embodied navigation. Unlike model adaptation during the training phase, test-time adaptation updates models to unlabeled data online during the testing phase [235, 236], enabling better knowledge acquisition and accuracy gains in unknown deployment environments. Gao et al. [237] proposed fast-slow test-time adaptation (FSTTA) for VLN tasks, balancing adaptability and stability by alternately performing fast updates for immediate adaptability and slow updates to mitigate cumulative errors during testing. Extensive experiments demonstrate that FSTTA significantly improves performance on four popular evaluation benchmarks, validating its effectiveness.

4.3.2 *Enhancing sensor reliability*

Sensor noise refers to the random variations and inaccuracies in the data collected by the sensors of embodied agents. These variations can arise from hardware limitations, environmental interference, and malicious attacks, significantly impacting the agent's perception of its environment, and leading to errors in navigation and decision-making.

Liu et al. [238] pioneered the study of vision adversarial attacks for embodied agents by generating spatiotemporal perturbations to create examples with 3D adversarial noises. They develop a trajectory attention module to examine scene view contributions, aiding in the localization of 3D objects that elicit the strongest stimuli for agents' predictions. By leveraging temporal clues and targeting spatial dimensions, they perturb the physical properties (e.g., 3D shape and texture) of key contextual objects in the most critical scene views. The results show that these 3D adversarial examples effectively attack state-of-the-art embodied agent models and significantly outperform other 3D adversarial attack methods.

Ying et al. [239] validated the malicious adversarial noises of embodied agents in vision navigation as well. They introduce δ-Markov decision process (δ-MDP) to model the persistent effect of universal adversarial perturbations (UAP) and propose two novel consistent attack methods, reward UAP and trajectory UAP. Extensive experiments reveal the significant reduction in the performance of existing embodied vision navigation methods under their proposed malicious adversarial attacks. Additionally, Tian et al. [240] investigated multisensory perception under adversarial attacks. They target audio-visual event recognition tasks as a proxy for audio-visual embodied navigation tasks to validate multimodal adversarial attacks. Concurrently, an audio-visual defense strategy is proposed based on an audio-visual dissimilarity constraint and external feature memory banks. Their experiments demonstrate that audio-visual models are vulnerable to multimodal adversarial attacks, and their defense method enhances the robustness of audio-visual models without significantly compromising clean model performance.

5 Embodied navigation enabled tasks

This section focuses on the tasks enabled by embodied navigation as in Figure 9, highlighting its indispensable role in the functionality and enhancement of advanced robotic systems. By demonstrating its transformative impact across various domains, we illustrate how embodied navigation is a critical

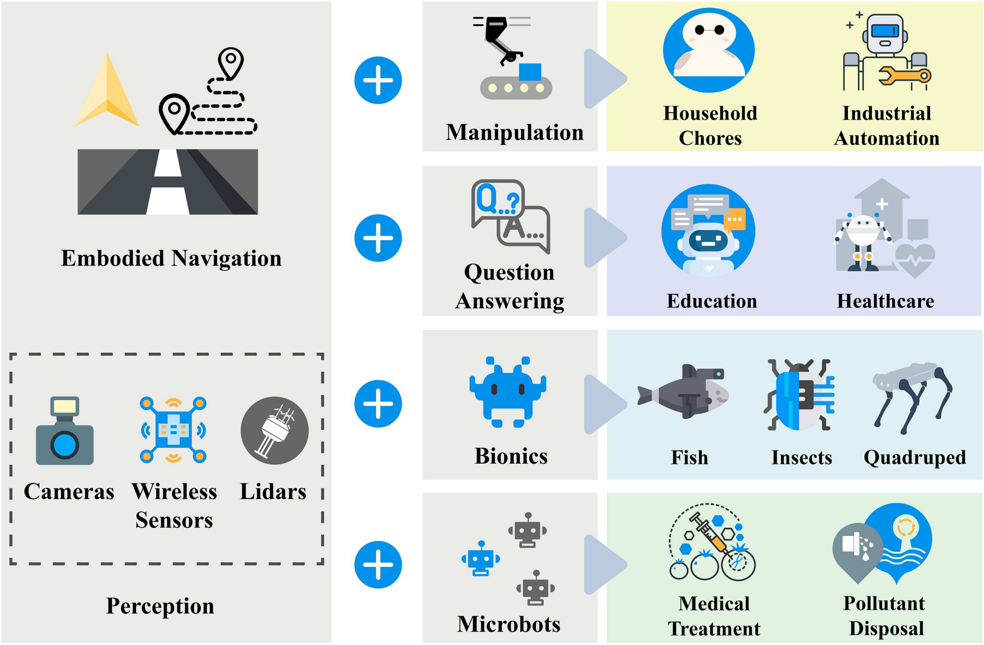

Figure 9   Tasks enabled by embodied navigation.

component that not only expands operational capabilities but also enables new possibilities in robotics and AI.

5.1 Autonomous driving

Embodied navigation's nature allows it to directly benefit autonomous driving platforms like cars and drones, enabling vehicles to perceive and respond to dynamic environments in real time [241–246]. Traditional systems often rely on predefined routes and static maps with limited adaptability. In contrast, embodied navigation allows these platforms to navigate safely and efficiently through unpredictable conditions and make instant decisions.

5.2 General assistant robot

General assistant robots rely on embodied navigation to function effectively in various environments, making them capable of performing tasks that would be impossible with static or limited mobility systems. These robots integrate navigation with manipulation and interaction capabilities, enabling them to operate autonomously in complex and unstructured settings.

5.2.1 *Navigation with manipulation*

Embodied navigation enables robots to navigate through complex environments to reach and manipulate objects. Without this capability, a robot's manipulation tasks would be confined to a limited area. Navigation broadens the scope of manipulation by allowing robots to move to different locations, adapt to new tasks, and interact with a variety of objects, thereby enhancing their utility in tasks like household chores and industrial automation.

For instance, TidyBot [9] combines these capabilities to automate household cleanup tasks by leveraging LLMs to infer generalized rules from user preferences. TidyBot's use of predefined manipulation primitives resulted in successfully placing 85% of objects in correct receptacles during real-world evaluations, demonstrating its efficacy in personalizing and executing household tasks. Similarly, OK-Robot [10] showcases the potential of VLMs integrated with robotics primitives for pick-and-drop operations in home environments. Utilizing pre-trained models for object detection and navigation, OK-Robot achieved a 58.5% success rate in zero-shot tasks, highlighting how open-knowledge models can enhance mobile manipulation without additional training. This approach underscores the versatility of combining navigation with manipulation to handle unstructured settings effectively.

To further ensure that LLM and VLM-generated policy instructions can be executed by robots in the current environment, SayCan [247] introduces a novel affordance function. This function combines LLM-generated instructions with environmental affordances to determine the robot's next executable action by scoring the feasibility of actions based on current conditions. This integration significantly enhances the flexibility and applicability of robotic navigation and manipulation, enabling robots to adapt to varying scenarios with greater precision. In a different yet complementary vein, Inner Monologue [248] integrates introspective reasoning into robotic tasks, where the robot maintains an internal dialogue to improve task planning and execution. This method enhances decision-making processes by allowing the robot to assess its actions and environmental conditions continually. By incorporating this form of reflective reasoning, robots achieve higher accuracy in manipulation tasks, particularly in dynamic and unpredictable settings.

Expanding on the flexibility of these integrations, the code-as-policies (CaP) framework [249] uses LLMs to generate policy code directly for robotic tasks, addressing the limitations of predefined skills. By interpreting natural language instructions and generating executable code, CaP enables robots to perform complex navigation and manipulation tasks. This method enhances the agent's capability to adapt to new tasks and environments with minimal data collection, which is essential for assistant robots.

5.2.2 *Navigation with question-answering*

The integration of navigation with question-answering systems allows robots to provide contextual and spatially relevant information. Embodied navigation enriches the dialogue by enabling robots to understand and incorporate spatial context into their responses, move to relevant locations, and perform actions based on questions asked. This capability is crucial for providing comprehensive assistance in healthcare and educational settings, where understanding and interacting with the environment significantly enhances the quality of service.

Das et al. [250] introduced the task of embodied question answering (EQA) where an AI agent must navigate a 3D environment to answer questions about its surroundings based on first-person vision. This foundational model highlights the integration of language understanding, visual recognition, and active perception, which are critical for the agent to navigate and gather necessary visual information efficiently. This research underscored the importance of embodied navigation as it directly affects the agent's ability to locate and interpret relevant visual cues to answer questions accurately.

Building on the challenges of navigation and interaction within EQA, Gordon et al. [11] introduced a more dynamic aspect to the task through their interactive question answering (IQA) model. Their work addresses the limitation of static environments in the initial EQA model by enabling the agent to manipulate objects within its environment, such as opening doors or rearranging items, to access previously unreachable areas or uncover hidden information. This enhancement allows the agent to perform tasks that require an understanding of both the environment's layout and the functional aspects of objects within it, significantly expanding the scope of questions the agent can answer. Further expanding the complexity of the navigational tasks, Yu et al. [251] introduced the multi-target embodied question answering (MT-EQA) model, which allows an agent to address questions involving multiple targets within the same environment. This development came in response to the limitations of single-target queries in earlier models. By enabling comparative reasoning across different spatial locations, the MT-EQA model allows the agent to perform more complex navigational tasks and answer a broader range of questions. This model not only tests the agent's navigational abilities but also its capacity to synthesize information from multiple locations and objects to formulate accurate responses.

PaLM-E [252] represents a significant advancement in the integration of LLMs with embodied navigation and question-answering, showcasing how multimodal inputs, including visual data and continuous sensor readings, can be seamlessly integrated into language models to enhance robotic decision-making and navigation capabilities.

5.3 Navigation for bionic

For bionic robots that mimic biological organisms, embodied navigation is a key component. These robots aim to replicate the mobility and perception of living beings, which is crucial for navigating complex terrains.

Embodied navigation, inspired by biological principles, enables bionic robots to move and interact like their natural counterparts, making them suitable for search and rescue, environmental monitoring, and exploration in diverse and challenging conditions.

5.3.1 *Bionic fish*

Embodied navigation endows bionic fish with enhanced underwater maneuverability and adaptability, essential for exploring and interacting with deep-sea environments. Research on bionic fish [253, 254], inspired by deep-sea creatures such as the snailfish, demonstrates how embodied navigation is crucial for underwater maneuverability. It developed a soft robotic fish capable of operating at extreme ocean depths, utilizing a distributed electronics system inspired by the snailfish's skull structure to enhance pressure resilience. This innovation addresses the challenges of deep-sea navigation and sampling without the need for bulky protective vessels. The future integration of advanced embodied navigation techniques could lead to even more autonomous, resilient, and efficient underwater exploration robots, capable of performing complex tasks such as deep-sea mining and environmental monitoring with minimal human intervention.

5.3.2 *Bionic insects*

Embodied navigation in bionic insects reduces reliance on complex external sensors, enabling streamlined design and efficient performance with minimal processing power. One study developed a wireless-controlled robotic insect [255], the BHMbot, which achieves ultrafast untethered running speeds. By mimicking the leg gaits of insects, this 2-cm-long microrobot can navigate complex trajectories and obstacles, demonstrating high mobility and efficient locomotion in confined spaces. This capability is crucial for applications such as remote inspection and environmental monitoring. Another study explored the sensory and locomotion mechanisms of insects to enhance robotic efficiency [256]. Inspired by the way flies use airflow pressure on their antennae for navigation, this research integrated low-resolution sensors into microbots to improve performance without adding significant power demands. This method enables the microbots to navigate and adapt to their environment more effectively. Future work could focus on enhancing the autonomy and robustness of these insect-scale robots, making them capable of performing intricate tasks such as infrastructure inspection, environmental monitoring, and precision agriculture.

5.3.3 *Multi-environment robots*

Embodied navigation enhances the versatility and operational efficiency of robots designed to navigate diverse terrains. A study on a quadruped robot [257] utilized dynamic control policies and state estimation to improve stability and adaptability in varied terrestrial environments. This approach leverages real-time feedback and embodied navigation to maintain stability across different terrains, highlighting the robot's ability to adapt to changing conditions. Conversely, a bio-inspired turtle robot [258] demonstrated the ability to navigate both land and water using adaptive morphology, allowing smooth transitions between swimming and walking. This design, inspired by turtle locomotion, showcases the potential of embodied navigation to enhance versatility and efficiency in multi-environment operations. Future developments could see amphibious robots equipped with more advanced sensory and adaptive systems, further enhancing their ability to tackle diverse terrains and perform a wide range of tasks from environmental monitoring to disaster response.

5.3.4 *Shape-morphing robots*

Embodied navigation allows shape-morphing robots to adapt their morphology to targeted routes, highlighting their potential in dynamic environments. Shape-morphing robots, inspired by creatures like octopuses and worms, showcase the advantages of embodied navigation in adapting to various tasks and environments. Researchers have developed shape-programmable soft robots [259, 260] that can change morphology to perform tasks such as grasping and locomotion, highlighting the role of embodied navigation in enabling the robot to interact dynamically with its surroundings. Tang introduces a worm-like soft robot designed for navigating sub-centimeter diameter pipelines [261], inspired by peristaltic motion. The robot uses dielectric elastomer actuators (DEAs) for propulsion and anchoring, allowing it to traverse various pipeline geometries including L-shaped, S-shaped, and spiral pipes. It demonstrates rapid motion and adaptability, critical for tasks like pipeline inspection. Future advancements could enhance the adaptability and multi-functionality of shape-morphing robots, making them invaluable in search and rescue operations, where these robots could navigate through debris and confined spaces to locate and assist survivors.

5.4 Navigation in micro-environment

This subsection explores the application of embodied navigation in micro-environments, focusing on miniaturized robots used for intricate tasks that require maneuvering through confined spaces, adapting to minute changes, and operating with high precision [262].

5.4.1 *Medical treatment*

Navigation-empowered micro-robots can be revolutionary to medical procedures by allowing for minimally invasive surgeries and targeted drug delivery and improving environmental monitoring by enabling detailed analysis in hard-to-reach areas [12]. Schmidt et al. [13] explored targeted drug delivery systems using microrobots, specifically designed for cancer treatment. It details how microrobots can be guided to tumor sites using external magnetic fields, providing a focused therapeutic approach that minimizes damage to surrounding healthy tissues. Dekanovsky et al. [14] discussed the use of microrobots in the targeted transport of hormones within the body. These microrobots are designed to navigate through complex vascular systems, guided by external magnetic fields and real-time imaging techniques. Future research could focus on integrating navigation with real-time tracking, adaptive systems, and advanced propulsion mechanisms. By addressing these areas, embodied navigation can unlock new potentials for microrobots in medical treatments, improving patient outcomes through minimally invasive, highly targeted therapeutic interventions.

5.4.2 *Water pollutant disposal*

Embodied navigation can be pivotal in enhancing the effectiveness of various micro- and nanorobot systems in the context of water pollutant disposal. One study introduced temperature-responsive magnetic nanorobots, which use magnetic propulsion and thermosensitive aggregation to efficiently remove arsenic and atrazine from water, demonstrating high recovery and reusability rates [263]. Another study developed light-driven nano/micromotors that utilize UV light for propulsion, significantly improving the capture and degradation of nanoplastics without requiring chemical fuels, thus providing an eco-friendly solution [264]. If further incorporated into these systems, embodied navigation could hopefully leverage the field of water purification by providing more precise control and enhanced adaptability in micro-environments [265].

6 Challenges and future work

As an emerging field of study, embodied navigation presents numerous unresolved issues and corresponding challenges in the current research. In this section, we highlight potential future directions for advancement, outlining key areas where further exploration and innovation may address existing limitations and enhance the field, as illustrated in Figure 10.

6.1 Embodied navigation in reality

Although a lot of studies and benchmarks in embodied navigation have been proposed in simulation environments [266], there remains a lack of more in-depth and broader exploration in large-scale real-world settings, causing a reality-performance gap under distribution shift [267]. This can be attributed to two main reasons: the relative scarcity of real-world data and the absence of comprehensive real-world evaluation benchmarks. (i) Challenges in data. Due to the high cost of data collection and manual labor in real-world environments, researchers often use simulated data for their model training and validation, which results in performance degradation in reality. To reduce this reality-performance gap, Sim2Real can be an effective approach for transferring knowledge learned in simulation to reality when real-world data is limited [268, 269] or totally inaccessible [270–274] during training. (ii) Challenges in evaluation benchmarks. Besides data, a real-world evaluation benchmark is also necessary and relatively lacking for sufficient model validation in real environments. Building on the success of simulated evaluation benchmarks [275, 276], creating real-world benchmarks for embodied AI tasks is meaningful but also requires substantial human efforts and financial resources. Researchers have already made a few efforts in this area [277–280]. Nevertheless, there is still a considerable distance to completely eliminate the performance gap between simulation and reality. This gap persists not only due to the more complex and

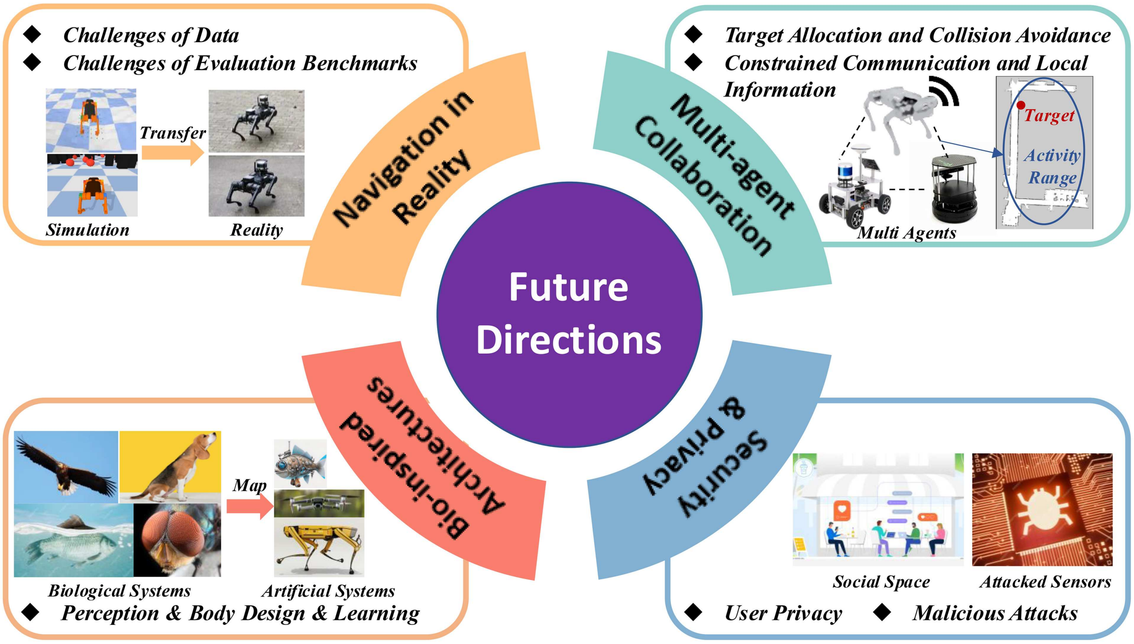

Figure 10 Challenges and future work of embodied navigation.

dynamic factors in the real environments but also because of the heterogeneous nature of real embodied agents, which have varying postures and behavior patterns, such as quadruped robots [281], modular robots [282, 283], and even biomimetic underwater robots [253]. Considering the behavior patterns of real embodied agents and dynamic task environments, constructing fair, scalable and effective real-world evaluation benchmarks will be a crucial long-term goal for future work.

6.2 Multi-agent collaboration

Multiple embodied agents can achieve swarm intelligence through collaboration that far surpasses the level of individual intelligence [284]. However, compared to single-agent systems [285, 286], current embodied navigation research pays relatively less attention to multi-agent systems. One reason is the lack of multi-agent features in existing evaluation benchmarks. To the best of our knowledge, only a few benchmarks [287–289] consider multi-agent setups, providing limited validation approaches for multi-agent algorithms. Moreover, unlike single-agent systems, multi-agent systems in embodied navigation encounter unique challenges. The agents must learn from their own observations and also account for the transitions of other agents, therefore requiring additional collaboration design in navigation policy, target allocation, and collision avoidance [290]. To address the above challenges, existing research develops collaboration algorithms via deep RL [291–293], GNN-based methods [294–296], or customized modular training [297–299].

Nevertheless, the challenges of multi-agent collaboration are far from being fully resolved. In more practical scenarios, multi-agent systems usually face communication constraints (e.g., bandwidth and range) [300]. Additionally, the collaboration of heterogeneous agents (i.e., agents with different functions and behavior patterns) has not been sufficiently explored, which is both meaningful and more challenging as it requires greater consideration of the complementary abilities of agents and a deeper scene understanding of the task [301]. Furthermore, compared to simply navigating to specific goals, multi-agent systems may need to undertake more complex formation tasks (e.g., flocking in a leader-follower configuration) without access to global information (i.e., a map and the states of all agents), where each agent has to navigate to an appropriate goal with minimal transitions, only knowing the states of nearby agents [302, 303]. In such cases, drawing inspiration from the collective behaviors of biological swarms (e.g., birds, fishes, and wolves) can be a promising and interesting research area [304, 305].

6.3 Bio-inspired neural architectures of embodied navigation

With the development of artificial neural networks and systems [306], embodied navigation has continuously achieved breakthroughs in accuracy and efficiency. However, we are still far from creating embodied agents that can interact naturally and safely with dynamic environments and each other, akin to biological organisms or humans. In nature, even small insects can handle different behavior patterns such as flying, swimming, and crawling, as well as robust navigation policies [307, 308]. Therefore, existing embodied navigation systems have yet to bridge the gap with biological neural systems.

The challenge lies in how to map biological neural systems to the perception, learning and behavior of embodied agents. Most existing embodied navigation studies focus on perception (e.g., multi-modal sensor feedbacks [179]) and learning (e.g., LLM-driven embodied policies [286]) and have achieved significant progress. On the other hand, research in robotic design [253, 309–311] focuses more on the "intelligence of embodied behaviors". These studies support the idea that, in many cases of biological organisms, intelligence can arise from the "embodiment" itself [312], such as the sensor distribution in insects [313] and the motion dynamics of animals [314]. Despite the above efforts, effective coordination of embodied agents' perception, learning, and behavior solutions remains unresolved. Possible approaches include (i) training individual solution modules and then assembling them into complete embodied policies [152], (ii) adopting end-to-end automatic training schemes [194], or (iii) further exploring and drawing inspiration from biological coordination mechanisms [315]. In the future exploration of validating these approaches, more support and breakthroughs in embodied hardware [316] and biological mechanisms [317] can also be indispensable.

6.4 Security and privacy

With the spread of embodied AI applications, embodied agents are increasingly playing a significant role in public spaces and human life. However, the complex functionalities of embodied agents (e.g., navigation, manipulation, and question-answering) and computing paradigms (e.g., edge computing) result in heightened uncertainties and increased susceptibility to diverse attacks on their sensors, computation, communication, and actuators, posing serious security risks and challenges. There exists a certain amount of studies exploring security problems for various types of embodied agents, such as social robots [318], UAVs [319], UGVs [320, 321], and high-level autonomous driving [322]. However, there has been relatively little discussion specifically addressing the security of embodied navigation [323]. To some extent, enhancing the robustness of embodied navigation (Subsection 4.3) can be seen as a branch of ensuring security.

Meanwhile, as an embodied agent is capable of collecting more private information from users in comparison to a disembodied one [324], user privacy problems (e.g., leakage of training data [325] and personal information [326, 327]) become increasingly severe in embodied AI. Recently, federated learning has been introduced to embodied navigation to provide a privacy-preserving embodied agent learning framework [328]. Certainly, it can be said that within the realm of embodied navigation, there persists a notable research gap regarding security and privacy concerns.

7 Concluding remark

In this article, we offer an in-depth review of embodied navigation research. We categorize the literature into four key areas: perception, navigation strategies, efficiency optimization, and embodied navigation enabled tasks, and delve into the critical topics within each category. We also highlight the substantial challenges that remain, including real-world applicability, multi-agent collaboration, bio-inspired neural architectures, and issues surrounding security and privacy. We aim for this article to serve as a foundational reference, providing researchers and practitioners with a thorough understanding of embodied navigation and inspiring further advancements in this promising and rapidly evolving field.

Acknowledgements This work was supported in part by National Key Research Plan of China (Grant No. 2021YFB2900100), National Natural Science Foundation of China (Grant Nos. 62202263, 62232004, 61925202), and Jiangsu Provincial Key Research and Development Program (Grant Nos. BE2022065-1, BE2022065-3). The authors would like to thank Hao WEN, Jinliang YUAN, Kebin LIU, Liangren XU, Xinqi JIN, Xu WANG for their contribution to this article.

References

1 Elfes A. Using occupancy grids for mobile robot perception and navigation. Computer, 1989, 22: 46–57

2 Shah D, Osiński B, Levine S, et al. LM-Nav: robotic navigation with large pre-trained models of language, vision, and action. In: Proceedings of Conference on Robot Learning, 2023. 492–504

3 Maturana D, Chou P W, Uenoyama M, et al. Real-time semantic mapping for autonomous off-road navigation. In: Proceedings of the 11th Conference on Field and Service Robotics, 2018. 335–350

4 Wen Z, Yang G, Cai Q, et al. A novel bluetooth-odometer-aided smartphone-based vehicular navigation in satellite-denied environments. IEEE Trans Ind Electron, 2023, 70: 3136–3146

5 Hu Z, Yuan J, Gao Y, et al. NALO-VOM: navigation-oriented LiDAR-guided monocular visual odometry and mapping for unmanned ground vehicles. IEEE Trans Intell Veh, 2024, 9: 2612–2623

6 Xu J, Cao H, Yang Z, et al. SwarmMap: scaling up real-time collaborative visual SLAM at the edge. In: Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22), 2022. 977–993

7 Chi G X, Yang Z, Xu J G, et al. Wi-Drone: Wi-Fi-based 6-DoF tracking for indoor drone flight control. In: Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, 2022. 56–68

8 Xu J G, Li D Y, Yang Z, et al. Taming event cameras with bio-inspired architecture and algorithm: a case for drone obstacle avoidance. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023

9 Wu J, Antonova R, Kan A, et al. TidyBot: personalized robot assistance with large language models. Auton Robot, 2023, 47: 1087–1102

10 Liu P Q, Orru Y, Vakil J, et al. OK-Robot: what really matters in integrating open-knowledge models for robotics. 2024. ArXiv:2401.12202

11 Gordon D, Kembhavi A, Rastegari M, et al. IQA: visual question answering in interactive environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4089–4098

12 Aziz A, Pane S, Iacovacci V, et al. Medical imaging of microrobots: toward in vivo applications. ACS Nano, 2020, 14: 10865–10893

13 Schmidt C K, Medina-Sánchez M, Edmondson R J, et al. Engineering microrobots for targeted cancer therapies from a medical perspective. Nat Commun, 2020, 11: 5618

14 Dekanovsky L, Khezri B, Rottnerova Z, et al. Chemically programmable microrobots weaving a web from hormones. Nat Mach Intell, 2020, 2: 711–718

15 Engel J, Stückler J, Cremers D. Large-scale direct SLAM with stereo cameras. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015. 1935–1942

16 Huang L Y. Review on LiDAR-based SLAM techniques. In: Proceedings of International Conference on Signal Processing and Machine Learning (CONF-SPML), 2021. 163–168

17 Khan M U, Zaidi S A A, Ishtiaq A, et al. A comparative survey of LiDAR-SLAM and LiDAR based sensor technologies. In: Proceedings of Mohammad Ali Jinnah University International Conference on Computing (MAJICC), 2021. 1–8

18 Franklin S. Autonomous agents as embodied AI. Cybern Syst, 1997, 28: 499–520

19 Lowe D G. Distinctive image features from scale-invariant keypoints. Int J Comput Vision, 2004, 60: 91–110

20 Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF). Comput Vision Image Understand, 2008, 110: 346–359

21 Rublee E, Rabaud V, Konolige K, et al. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of International Conference on Computer Vision, 2011. 2564–2571

22 Zhang Y D, Funkhouser T. Deep depth completion of a single RGB-D image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

23 Hirschmuller H, Scharstein D. Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans Pattern Anal Mach Intell, 2009, 31: 1582–1599

24 Meilland X, Comport A I, Rives P. Real-time dense visual tracking under large lighting variations. In: Proceedings of the British Machine Vision Conference, 2011

25 Gonçalves T, Comport A I. Real-time direct tracking of color images in the presence of illumination variation. In: Proceedings of IEEE International Conference on Robotics and Automation, 2011

26 Klose S, Heise P, Knoll A. Efficient compositional approaches for real-time robust direct visual odometry from RGB-D data. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013. 1100–1106

27 Scandaroli G G, Meilland M, Richa R. Improving NCC-based direct visual tracking. In: Proceedings of European Conference on Computer Vision, 2012. 442–455

28 Dame A, Marchand E. Second-order optimization of mutual information for real-time image registration. IEEE Trans Image Process, 2012, 21: 4190–4203

29 Crivellaro A, Lepetit V. Robust 3D tracking with descriptor fields. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014. 3414–3421

30 Dai A, Nießner M, Zollhöfer M, et al. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface re-integration. ACM Trans Graph, 2017, 36: 1

31 Zabih R, Woodfill J. Non-parametric local transforms for computing visual correspondence. In: Proceedings of European Conference on Computer Vision, 1994

32 Dong E Q, Xu J G, Wu C S, et al. Pair-Navi: peer-to-peer indoor navigation with mobile visual SLAM. In: Proceedings of IEEE Conference on Computer Communications, 2019. 1189–1197

33 Xu J G, Cao H, Li D Y, et al. Edge assisted mobile semantic visual SLAM. In: Proceedings of IEEE Conference on Computer Communications, 2020. 1828–1837

34 von Stumberg L, Wenzel P, Khan Q, et al. GN-Net: the Gauss-Newton loss for multi-weather relocalization. IEEE Robot Autom Lett, 2020, 5: 890–897

35 Sarlin P E, Unagar A, Larsson M, et al. Back to the feature: learning robust camera localization from pixels to pose. In: Proceedings of Computer Vision and Pattern Recognition, 2021

36 Lee D, Jung M, Yang W, et al. LiDAR odometry survey: recent advancements and remaining challenges. Intel Serv Robot, 2024, 17: 95–118

37 Besl P J, McKay N D. Method for registration of 3-D shapes. In: Proceedings of Sensor Fusion IV: Control Paradigms and Data Structures, 1992. 586–606

38 Cao Q, Liao Y, Fu Z, et al. An iterative closest point method for LiDAR odometry with fused semantic features. Appl Sci, 2023, 13: 12741

39 Zhang J, Singh S. LOAM: LiDAR odometry and mapping in real-time. In: Proceedings of Robotics: Science and Systems, 2014. 1–9

40 Zhang J, Singh S. Low-drift and real-time lidar odometry and mapping. Auton Robot, 2017, 41: 401–416

41 Shan T X, Englot B. LeGO-LOAM: lightweight and ground-optimized LiDAR odometry and mapping on variable terrain. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018. 4758–4765

42 Oelsch M, Karimi M, Steinbach E. R-LOAM: improving LiDAR odometry and mapping with point-to-mesh features of a known 3D reference object. IEEE Robot Autom Lett, 2021, 6: 2068–2075

43 Li Z C, Wang N Y. DMLO: deep matching LiDAR odometry. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 6010–6017

44 Zheng C, Lyu Y C, Li M, et al. LodoNet: a deep neural network with 2D keypoint matching for 3D LiDAR odometry estimation. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 2391–2399

45 Wang G, Wu X, Jiang S, et al. Efficient 3D deep LiDAR odometry. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 5749–5765

46 Zhang D, Peng T, Loomis J S. Optimized deep learning for LiDAR and visual odometry fusion in autonomous driving. IEEE Sens J, 2023, 23: 29594–29604

47 Li Q, Chen S Y, Wang C, et al. LO-Net: deep real-time LiDAR odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 8473–8482

48 Liu T, Wang Y, Niu X, et al. LiDAR odometry by deep learning-based feature points with two-step pose estimation. Remote Sens, 2022, 14: 2764

49 Abu-Alrub N J, Rawashdeh N A. Radar odometry for autonomous ground vehicles: a survey of methods and datasets. IEEE Trans Intell Veh, 2024, 9: 4275–4291

50 Amjad B, Ahmed Q Z, Lazaridis P I, et al. Radio SLAM: a review on radio-based simultaneous localization and mapping. IEEE Access, 2023, 11: 9260–9278

51 Liu R, Marakkalage S H, Padmal M, et al. Collaborative SLAM based on WiFi fingerprint similarity and motion information. IEEE Int Things J, 2020, 7: 1826–1840

52 Arun A, Ayyalasomayajula R, Hunter W, et al. P2SLAM: bearing based WiFi SLAM for indoor robots. IEEE Robot Autom Lett, 2022, 7: 3326–3333

53 Arun A, Hunter W, Ayyalasomayajula R, et al. WAIS: leveraging WiFi for resource-efficient SLAM. In: Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services, 2024. 561–574

54 Wang C, Zhang H D, Nguyen T M, et al. Ultra-wideband aided fast localization and mapping system. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017. 1602–1609

55 Cao Z Q, Liu R, Yuen C, et al. Relative localization of mobile robots with multiple ultra-wideband ranging measurements. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021. 5857–5863

56 Palacios J, Bielsa G, Casari P, et al. Communication-driven localization and mapping for millimeter wave networks. In: Proceedings of IEEE Conference on Computer Communications, 2018. 2402–2410

57 He J, Yin F, So H C. A framework for millimeter-wave multi-user SLAM and its low-cost realization. Signal Process, 2023, 209: 109018

58 Wu C, Gong Z, Tao B, et al. RF-SLAM: UHF-RFID based simultaneous tags mapping and robot localization algorithm for smart warehouse position service. IEEE Trans Ind Inf, 2023, 19: 11765–11775

59 Fu G, Zhang J, Chen W, et al. Precise localization of mobile robots via odometry and wireless sensor network. Int J Adv Robot Syst, 2013, 10: 203

60 Mirowski P, Ho T K, Yi S, et al. SignalSLAM: simultaneous localization and mapping with mixed WiFi, Bluetooth, LTE and magnetic signals. In: Proceedings of International Conference on Indoor Positioning and Indoor Navigation, 2013. 1–10

61 Cai G S, Lin H Y, Kao S F. Mobile robot localization using GPS, IMU and visual odometry. In: Proceedings of International Automatic Control Conference (CACS), 2019. 1–6

62 Li D, Zhang F, Feng J, et al. LD-SLAM: a robust and accurate GNSS-aided multi-map method for long-distance visual SLAM. Remote Sens, 2023, 15: 4442

63 Liu W, Caruso D, Ilg E, et al. TLIO: tight learned inertial odometry. IEEE Robot Autom Lett, 2020, 5: 5653–5660

64 Chen C, Lu C X, Wahlstrom J, et al. Deep neural network based inertial odometry using low-cost inertial measurement units. IEEE Trans Mobile Comput, 2021, 20: 1351–1364

65 Liu H, Wei X, Perusquía-Hernández M, et al. DUET: improving inertial-based odometry via deep IMU online calibration. IEEE Trans Instrum Meas, 2023, 72: 1–13

66 Chen C, Pan X. Deep learning for inertial positioning: a survey. IEEE Trans Intell Transp Syst, 2024, 25: 10506–10523

67 Vödisch N, Cattaneo D, Burgard W, et al. CoVIO: online continual learning for visual-inertial odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 2464–2473

68 Gui L, Zeng C, Dauchert S, et al. A ZUPT aided initialization procedure for tightly-coupled lidar inertial odometry based SLAM system. J Intell Robot Syst, 2023, 108: 40

69 Xu X, Zhang L, Yang J, et al. A review of multi-sensor fusion SLAM systems based on 3D LIDAR. Remote Sens, 2022, 14: 2835

70 Chen W, Zhou C, Shang G, et al. SLAM overview: from single sensor to heterogeneous fusion. Remote Sens, 2022, 14: 6033

71 Yang M, Sun X, Jia F, et al. Sensors and sensor fusion methodologies for indoor odometry: a review. Polymers, 2022, 14: 2019

72 Zhu J, Li H, Zhang T. Camera, LiDAR, and IMU based multi-sensor fusion SLAM: a survey. Tsinghua Sci Technol, 2024, 29: 415–429

73 Graeter J, Wilczynski A, Lauer M. LIMO: LiDAR-monocular visual odometry. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018. 7872–7879

74 Wang W, Liu J, Wang C, et al. DV-LOAM: direct visual LiDAR odometry and mapping. Remote Sens, 2021, 13: 3340

75 Yuan Z, Cheng J, Yang X. CR-LDSO: direct sparse LiDAR-assisted visual odometry with cloud reusing. IEEE Trans Multimedia, 2023, 25: 9397–9409

76 Zhang J, Singh S. Visual-LiDAR odometry and mapping: low-drift, robust, and fast. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2015. 2174–2181

77 Huang S S, Ma Z Y, Mu T J, et al. LiDAR-monocular visual odometry using point and line features. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2020. 1091–1097

78 Huang K H, Xiao J H, Stachniss C. Accurate direct visual-laser odometry with explicit occlusion handling and plane detection. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2019. 1295–1301

79 Stumberg L, Cremers D. DM-VIO: delayed marginalization visual-inertial odometry. IEEE Robot Autom Lett, 2022, 7: 1408–1415

80 Li J, Pan X, Huang G, et al. RD-VIO: robust visual-inertial odometry for mobile augmented reality in dynamic environments. IEEE Trans Visual Comput Graph, 2024, 30: 6941–6955

81 Chen P, Guan W, Lu P. ESVIO: event-based stereo visual inertial odometry. IEEE Robot Autom Lett, 2023, 8: 3661–3668

82 Xu W, Cai Y, He D, et al. FAST-LIO2: fast direct LiDAR-inertial odometry. IEEE Trans Robot, 2022, 38: 2053–2073

83 Kim B, Jung C, Shim D H, et al. Adaptive keyframe generation based LiDAR inertial odometry for complex underground environments. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2023. 3332–3338

84 Lim H, Kim D, Kim B, et al. AdaLIO: robust adaptive LiDAR-inertial odometry in degenerate indoor environments. In: Proceedings of the 20th International Conference on Ubiquitous Robots (UR), 2023. 48–53

85 Shan T X, Englot B, Ratti C, et al. LVI-SAM: tightly-coupled LiDAR-visual-inertial odometry via smoothing and mapping. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2021. 5692–5698

86 Zhao Z, Zhang Y, Long L, et al. Efficient and adaptive LiDAR-visual-inertial odometry for agricultural unmanned ground vehicle. Int J Adv Robotic Syst, 2022, 19: 17298806221094925

87 Zhang H, Du L, Bao S, et al. LVIO-Fusion: tightly-coupled LiDAR-visual-inertial odometry and mapping in degenerate environments. IEEE Robot Autom Lett, 2024, 9: 3783–3790

88 Newman P, Ho K. SLAM-loop closing with visually salient features. In: Proceedings of the IEEE International Conference on Robotics and Automation, 2005. 635–642

89 Cummins M, Newman P. FAB-MAP: probabilistic localization and mapping in the space of appearance. Int J Robot Res, 2008, 27: 647–665

90 Angeli A, Doncieux S, Meyer J A, et al. Incremental vision-based topological SLAM. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. 1031–1036

91 Murillo A C, Kosecka J. Experiments in place recognition using gist panoramas. In: Proceedings of the 12th International Conference on Computer Vision Workshops, 2009. 2196–2203

92 Sünderhauf N, Protzel P. Brief-gist — closing the loop by simple means. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011. 1234–1241

93 Samadzadeh A, Nickabadi A. SRVIO: super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions. IEEE Trans Robot, 2023, 39: 2878–2891

94 Arshad S, Kim G W. Role of deep learning in loop closure detection for visual and LiDAR SLAM: a survey. Sensors, 2021, 21: 1243

95 Magnusson M, Andreasson H, Nüchter A, et al. Automatic appearance-based loop detection from three-dimensional laser data using the normal distributions transform. J Field Robot, 2009, 26: 892–914

96 Magnusson M, Andreasson H, Nuchter A, et al. Appearance-based loop detection from 3D laser data using the normal distributions transform. In: Proceedings of IEEE International Conference on Robotics and Automation, 2009. 23–28

97 Bosse M, Zlot R. Keypoint design and evaluation for place recognition in 2D LiDAR maps. Robot Auton Syst, 2009, 57: 1211–1224

98 Dubé R, Dugas D, Stumm E, et al. SegMatch: segment based place recognition in 3D point clouds. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2017. 5266–5272

99 Liu X, Zhang L, Qin S, et al. Optimized LOAM using ground plane constraints and SegMatch-based loop detection. Sensors, 2019, 19: 5419

100 Zhou B, Li C, Chen S, et al. ASL-SLAM: a LiDAR SLAM with activity semantics-based loop closure. IEEE Sens J, 2023, 23: 13499–13510

101 Arce J, Vodisch N, Cattaneo D, et al. PADLoC: LiDAR-based deep loop closure detection and registration using panoptic attention. IEEE Robot Autom Lett, 2023, 8: 1319–1326

102 Checchin P, Gérossier F, Blanc C, et al. Radar scan matching SLAM using the Fourier-Mellin transform. In: Springer Tracts in Advanced Robotics. Berlin: Springer, 2010. 151–161

103 Hong Z Y, Petillot Y, Wang S. RadarSLAM: radar based large-scale SLAM in all weathers. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 5164–5170

104 Gao J, Fan J, Zhai S, et al. Wi-Loop SLAM: loop closures with wireless sensing in multipath SLAM. IEEE Trans Wireless Commun, 2024, 23: 15185–15197

105 Adolfsson D, Karlsson M, Kubelka V, et al. TBV radar SLAM — trust but verify loop candidates. IEEE Robot Autom Lett, 2023, 8: 3613–3620

106 Kim G, Kim A. Scan context: egocentric spatial descriptor for place recognition within 3D point cloud map. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018. 4802–4809

107 Kim G, Park Y S, Cho Y, et al. MulRan: multimodal range dataset for urban place recognition. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2020. 6246–6253

108 Wang Y, Ma H. mVIL-Fusion: monocular visual-inertial-LiDAR simultaneous localization and mapping in challenging environments. IEEE Robot Autom Lett, 2023, 8: 504–511

109 Liu Z, Li Z, Liu A, et al. LVI-Fusion: a robust LiDAR-visual-inertial SLAM scheme. Remote Sens, 2024, 16: 1524

110 Zhao X, Wen C, Prakhya S M, et al. Multimodal features and accurate place recognition with robust optimization for LiDAR-visual-inertial SLAM. IEEE Trans Instrum Meas, 2024, 73: 1–16

111 Pan H, Liu D, Ren J, et al. LiDAR-IMU tightly-coupled SLAM method based on IEKF and loop closure detection. IEEE J Sel Top Appl Earth Obs Remote Sens, 2024, 17: 6986–7001

112 Qin T, Li P, Shen S. VINS-Mono: a robust and versatile monocular visual-inertial state estimator. IEEE Trans Robot, 2018, 34: 1004–1020

113 Galvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences. IEEE Trans Robot, 2012, 28: 1188–1197

114 Shan T X, Englot B, Meyers D, et al. LIO-SAM: tightly-coupled LiDAR inertial odometry via smoothing and mapping. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 5135–5142

115 Chghaf M, Flórez S R, Ouardi A E. A multimodal loop closure fusion for autonomous vehicles SLAM. Robot Auton Syst, 2023, 165: 104446

116 Jiang F, Wang W, You H, et al. TS-LCD: two-stage loop-closure detection based on heterogeneous data fusion. Sensors, 2024, 24: 3702

117 Kalman R E. A new approach to linear filtering and prediction problems. J Basic Eng, 1960, 82: 35–45

118 Urrea C, Agramonte R, Diraco G. Kalman filter: historical overview and review of its use in robotics 60 years after its creation. J Sens, 2021, 2021: 9674015

119 Schneider T, Dymczyk M, Fehr M, et al. Maplab: an open framework for research in visual-inertial mapping and localization. IEEE Robot Autom Lett, 2018, 3: 1418–1425

120 Cramariuc A, Bernreiter L, Tschopp F, et al. Maplab 2.0 — a modular and multi-modal mapping framework. IEEE Robot Autom Lett, 2023, 8: 520–527

121 Campos C, Elvira R, Rodriguez J J G, et al. ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Trans Robot, 2021, 37: 1874–1890

122 Klingner B, Martin D, Roseborough J. Street view motion-from-structure-from-motion. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013

123 Heinly J, Schonberger J L, Dunn E, et al. Reconstructing the world* in six days *(as captured by the Yahoo 100 million image dataset). In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

124 Thomee B, Shamma D A, Friedland G, et al. YFCC100M: the new data in multimedia research. Commun ACM, 2016, 59:

64–73
125 Zhanabatyrova A, Leite C S, Xiao Y. Structure from motion-based mapping for autonomous driving: practice and experience. ACM Trans Int Things, 2024, 5: 1–25
126 Google. How navigation data makes Maps better for everyone. 2024. https://support.google.com/maps/answer/10565726
127 Hosseini M, Timmerer C. Dynamic adaptive point cloud streaming. In: Proceedings of the 23rd Packet Video Workshop, 2018. 25–30
128 van der Hooft J, Wauters T, De Turck F, et al. Towards 6DoF HTTP adaptive streaming through point cloud compression. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019. 2405–2413
129 Lee K, Yi J, Lee Y, et al. GROOT: a real-time streaming system of high-fidelity volumetric videos. In: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, 2020
130 Yang Z, Zhou Z, Liu Y. From RSSI to CSI: indoor localization via channel response. ACM Comput Surv, 2013, 46: 1–32
131 Yang Z, Wu C S, Liu Y H. Locating in fingerprint space: wireless indoor localization with little human intervention. In: Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, 2012. 269–280
132 Wu C S, Yang Z, Liu Y H, et al. WILL: wireless indoor localization without site survey. IEEE Trans Parallel Distrib Syst, 2013, 24: 839–848
133 Ding Y, Liu L, Yang Y, et al. From conception to retirement: a lifetime story of a 3-year-old wireless beacon system in the wild. In: Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), 2021. 859–872
134 Ding Y, Yang Y, Jiang W C, et al. Nationwide deployment and operation of a virtual arrival detection system in the wild. In: Proceedings of the ACM SIGCOMM Conference, 2021. 705–717
135 Siam S I, Ahn H, Liu L, et al. Artificial intelligence of things: a survey. ACM Trans Sen Netw, 2025, 21: 1–75
136 Geok T K, Aung K Z, Aung M S, et al. Review of indoor positioning: radio wave technology. Appl Sci, 2021, 11: 279
137 Liu F, Liu J, Yin Y, et al. Survey on WiFi-based indoor positioning techniques. IET Commun, 2020, 14: 1372–1383
138 Apple Inc. iBeacon — Apple Developer. 2024. https://developer.apple.com/ibeacon/
139 Baidu. Baidu In-Vehicle Navigation Map. 2024. https://en.apollo.auto/mapauto
140 Yin H, Xu X, Lu S, et al. A survey on global LiDAR localization: challenges, advances and open problems. Int J Comput Vis, 2024, 132: 3139–3171
141 Sarlin P E, Cadena C, Siegwart R, et al. From coarse to fine: robust hierarchical localization at large scale. In: Proceedings of Computer Vision and Pattern Recognition, 2019
142 Shotton J, Glocker B, Zach C, et al. Scene coordinate regression forests for camera relocalization in RGB-D images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013. 2930–2937
143 Guzman-Rivera A, Kohli P, Glocker B, et al. Multi-output learning for camera relocalization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014. 1114–1121
144 Valentin J, Nießner M, Shotton J, et al. Exploiting uncertainty in regression forests for accurate camera relocalization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 4400–4408
145 Brachmann E, Rother C. Learning less is more — 6D camera localization via 3D surface regression. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 4654–4662
146 Yang L W, Bai Z Q, Tang C Z, et al. SANet: scene agnostic network for camera localization. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2019. 42–51
147 Chen S, Cavallari T, Prisacariu V A, et al. Map-relative pose regression for visual re-localization. In: Proceedings of Computer Vision and Pattern Recognition, 2024
148 amap_tech. How AMAP's Improving Positioning Precision. 2024. https://www.alibabacloud.com/blog/how-amaps-improving-positioning-precision_596143
149 Wu K S, Xiao J, Yi Y W, et al. FILA: fine-grained indoor localization. In: Proceedings of 2012 Proceedings IEEE INFOCOM, 2012. 2210–2218
150 Xiong J, Jamieson K. ArrayTrack: a fine-grained indoor location system. In: Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13), 2013. 71–84
151 Soltanaghaei E, Kalyanaraman A, Whitehouse K. Multipath triangulation: decimeter-level WiFi localization and orientation with a single unaided receiver. In: Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, 2018. 376–388
152 Gan C, Zhang Y W, Wu J J, et al. Look, listen, and act: towards audio-visual embodied navigation. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2020. 9701–9707
153 Chen C G, Jain U, Schissler C, et al. SoundSpaces: audio-visual navigation in 3D environments. In: Proceedings of European Conference on Computer Vision, 2020
154 Chen J Y, Wang W G, Liu S, et al. Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. 10993–11003
155 Weinstein R. RFID: a technical overview and its application to the enterprise. IT Prof, 2005, 7: 27–33
156 Ni L M, Liu Y H, Lau Y C, et al. LANDMARC: indoor location sensing using active RFID. In: Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications, 2003. 407–415
157 Yang L, Chen Y K, Li X Y, et al. Tagoram: real-time tracking of mobile RFID tags to high precision using cots devices. In: Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, 2014. 237–248
158 Zhao Y Y, Liu Y H, Ni L M. VIRE: active RFID-based localization using virtual reference elimination. In: Proceedings of International Conference on Parallel Processing (ICPP 2007), 2007. 56–56
159 RAIN Alliance. 2023 RAIN RFID TAG IC Shipment Data. 2024. https://rainrfid.org/2023-rain-rfid-tag-ic-shipment-data/
160 Beontag. Logistcis and Supply Chain. 2024. https://www.beontag.com/rfid/application/logistics-and-supply-chain/
161 Shangguan L F, Jamieson K. The design and implementation of a mobile RFID tag sorting robot. In: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, 2016. 31–42
162 Boroushaki T, Leng J S, Clester I, et al. Robotic grasping of fully-occluded objects using RF perception. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2021. 923–929
163 Boroushaki T, Perper I, Nachin M, et al. RFusion: robotic grasping via RF-visual sensing and learning. In: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, 2021. 192–205
164 Boroushaki T, Dodds L, Naeem N, et al. FuseBot: mechanical search of rigid and deformable objects via multi-modal perception. Auton Robot, 2023, 47: 1137–1154
165 Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware CNN model. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 1134–1142
166 He K M, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2961–2969
167 Chaplot D S, Gandhi D P, Gupta A, et al. Object goal navigation using goal-oriented semantic exploration. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020. 4247–4258
168 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020. ArXiv:2010.11929

169 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of European Conference on Computer Vision, 2020. 213–229

170 Zhang J M, Yang K L, Constantinescu A, et al. Trans4Trans: efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 1760–1770

171 Wu Y, Wu Y X, Tamar A, et al. Bayesian relational memory for semantic visual navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 2769–2779

172 Mousavian A, Toshev A, Fišer M, et al. Visual representations for semantic target driven navigation. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2019. 8846–8852

173 Yang W, Wang X L, Farhadi A, et al. Visual semantic navigation using scene priors. 2018. ArXiv:1810.06543

174 Du H M, Yu X, Zheng L. VTNet: visual transformer network for object goal navigation. 2021. ArXiv:2105.09447

175 Hao W T, Li C Y, Li X J, et al. Towards learning a generic agent for vision-and-language navigation via pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 13137–13146

176 Pashevich A, Schmid C, Sun C. Episodic transformer for vision-and-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 15942–15952

177 Suglia A, Gao Q Z, Thomason J, et al. Embodied BERT: a transformer model for embodied, language-guided visual task completion. 2021. ArXiv:2108.04927

178 Chen C G, Al-Halah Z, Grauman K. Semantic audio-visual navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 15516–15525

179 Paul S, Roy-Chowdhury A, Cherian A. AVLEN: audio-visual-language embodied navigation in 3D environments. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022. 6236–6249

180 Yan R, Yang K L, Wang K W. NLFNet: non-local fusion towards generalized multimodal semantic segmentation across RGB-depth, polarization, and thermal images. In: Proceedings of IEEE International Conference on Robotics and Biomimetics (ROBIO), 2021. 1129–1135

181 Hart P, Nilsson N, Raphael B. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cyber, 1968, 4: 100–107

182 Stentz A. Optimal and efficient path planning for partially-known environments. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1994. 3310–3317

183 Leonard J J, Durrant-Whyte H F. Simultaneous map building and localization for an autonomous mobile robot. In: Proceedings of IEEE/RSJ International Workshop on Intelligent Robots and Systems'91, 1991. 1442–1447

184 Bongard J. Probabilistic robotics. sebastian thrun, wolfram burgard, and dieter fox. Artif Life, 2008, 14: 227–229

185 Luo H K, Yue A, Hong Z W, et al. Stubborn: a strong baseline for indoor object navigation. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. 3287–3293

186 Ramakrishnan S K, Chaplot D S, Al-Halah Z, et al. PONI: potential functions for objectgoal navigation with interaction-free learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 18890–18900

187 Georgakis G, Bucher B, Schmeckpeper K, et al. Learning to map for active semantic goal navigation. 2021. ArXiv:2106.15648

188 Zhu M Z, Zhao B L, Kong T. Navigating to objects in unseen environments by distance prediction. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. 10571–10578

189 Kuipers B. Modeling spatial knowledge. Cogn Sci, 1978, 2: 129–153

190 Thrun S, Bücken A. Integrating grid-based and topological maps for mobile robot navigation. In: Proceedings of the National Conference on Artificial Intelligence, 1996. 944–951

191 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2016. ArXiv:1609.02907

192 Kiran D A, Anand K, Kharyal C, et al. Spatial relation graph and graph convolutional network for object goal navigation. In: Proceedings of the 18th International Conference on Automation Science and Engineering (CASE), 2022, 1392–1398

193 Liu J J, Guo J F, Meng Z H, et al. ReVoLT: relational reasoning and Voronoi local graph planning for target-driven navigation. 2023. ArXiv:2301.02382

194 Ye J, Batra D, Das A, et al. Auxiliary tasks and exploration enable objectgoal navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 16117–16126

195 Ramrakhya R, Undersander E, Batra D, et al. Habitat-Web: learning embodied object-search strategies from human demonstrations at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 5173–5183

196 Al-Halah A, Ramakrishnan S K, Grauman K. Zero experience required: Plug&Play modular transfer learning for semantic visual navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 17031–17041

197 Ramrakhya R, Batra D, Wijmans E, et al. PIRLNav: pretraining with imitation and RL finetuning for ObjectNav. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 17896–17906

198 Zhang J Z, Dai L, Meng F P, et al. 3D-Aware object goal navigation via simultaneous exploration and identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 6672–6682

199 Singh K P, Weihs L, Herrasti A, et al. Ask4Help: learning to leverage an expert for embodied tasks. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022. 16221–16232

200 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763

201 Jia C, Yang Y F, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of International Conference on Machine Learning, 2021. 4904–4916

202 Singh A, Hu R H, Goswami V, et al. FLAVA: a foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 15638–15650

203 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020. 1877–1901

204 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res, 2020, 21: 1–67

205 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805

206 Majumdar A, Aggarwal G, Devnani B, et al. ZSON: zero-shot object-goal navigation using multimodal goal embeddings. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022. 32340–32352

207 Zhou K W, Zheng K Z, Pryor C, et al. ESC: exploration with soft commonsense constraints for zero-shot object navigation. In: Proceedings of International Conference on Machine Learning, 2023, 42829–42842

208 Zhou G Z, Hong Y C, Wu Q. NavGPT: explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 7641–7649

209 Schumann R, Zhu W R, Feng W X, et al. VELMA: verbalization embodiment of LLM agents for vision and language navigation in street view. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 18924–18933

210 Huang C G, Mees O, Zeng A, et al. Visual language maps for robot navigation. In: Proceedings of IEEE International

Conference on Robotics and Automation (ICRA), 2023. 10608–10615

211 LiKamWa R, Zhong L. Starfish: efficient concurrency support for computer vision applications. In: Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, 2015. 213–226

212 Han S, Shen H C, Philipose M, et al. MCDNN: an approximation-based execution framework for deep stream processing under resource constraints. In: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, 2016. 123–136

213 Fang B Y, Zeng X, Zhang M. NestDNN: resource-aware multi-tenant on-device deep learning for continuous mobile vision. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, MobiCom'18, 2018. 115–127

214 Wen H, Li Y C, Zhang Z S, et al. AdaptiveNet: post-deployment neural architecture adaptation for diverse edge environments. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023

215 Han R, Zhang Q L, Liu C H, et al. LegoDNN: block-grained scaling of deep neural networks for mobile vision. In: Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, 2021. 406–419

216 Feng X, Jiang Y, Yang X, et al. Computer vision algorithms and hardware implementations: a survey. Integration, 2019, 69: 309–320

217 Stone J E, Gohara D, Shi G. OpenCL: a parallel programming standard for heterogeneous computing systems. Comput Sci Eng, 2010, 12: 66–73

218 Liu S C, Zhou W T, Zhou Z M, et al. Deep learning inference on heterogeneous mobile processors: potentials and pitfalls. In: Proceedings of the Workshop on Adaptive AIoT Systems, 2024. 1–6

219 Wei J Y, Cao T, Cao S J, et al. NN-Stretch: automatic neural network branching for parallel inference on heterogeneous multi-processors. In: Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services, 2023. 70–83

220 Jeong J S, Lee J, Kim D, et al. Band: coordinated multi-DNN inference on heterogeneous mobile processors. In: Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, 2022. 235–247

221 Zhang C Y, Zhang F, Chen K Y, et al. EdgeNN: efficient neural network inference for CPU-GPU integrated edge devices. In: Proceedings of the 39th International Conference on Data Engineering (ICDE), 2023. 1193–1207

222 Niu W, Guan J X, Wang Y Z, et al. DNNFusion: accelerating deep neural networks execution with advanced operator fusion. In: Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, 2021. 883–898

223 Kimura N, Latifi S. A survey on data compression in wireless sensor networks. In: Proceedings of International Conference on Information Technology: Coding and Computing, 2005. 8–13

224 Arroyo-Valles R, Marques A G, Cid-Sueiro J. Optimal selective transmission under energy constraints in sensor networks. IEEE Trans Mobile Comput, 2009, 8: 1524–1538

225 Jiang Z H, Ling N W, Huang X, et al. CoEdge: a cooperative edge system for distributed real-time deep learning tasks. In: Proceedings of the 22nd International Conference on Information Processing in Sensor Networks, 2023. 53–66

226 Cui J H, Shi S Y, He Y Z, et al. VILAM: infrastructure-assisted 3D visual localization and mapping for autonomous driving. In: Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation, 2024. 1831–1845

227 Saeed R A, Ali E S, Abdelhaq M, et al. Energy efficient path planning scheme for unmanned aerial vehicle using hybrid generic algorithm-based Q-learning optimization. IEEE Access, 2024, 12: 13400–13417

228 Wang Z, Rong H, Jiang H, et al. A load-balanced and energy-efficient navigation scheme for UAV-mounted mobile edge computing. IEEE Trans Netw Sci Eng, 2022, 9: 3659–3674

229 Wei M H, Isler V. Energy-efficient path planning for ground robots by and combining air and ground measurements. In: Proceedings of the Conference on Robot Learning, 2020. 766–775

230 Åkerblom N, Chen Y X, Chehreghani M H. An online learning framework for energy-efficient navigation of electric vehicles. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2021

231 Modares J, Ghanei F, Mastronarde N, et al. UB-ANC planner: energy efficient coverage path planning with multiple drones. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2017. 6182–6189

232 Chattopadhyay P, Hoffman J, Mottaghi R, et al. RobustNav: towards benchmarking robustness in embodied navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 15691–15700

233 Hansen N, Jangir R, Sun Y, et al. Self-supervised policy adaptation during deployment. 2020. ArXiv:2007.04309

234 Peng J, Xu Y, Luo L, et al. Regularized denoising masked visual pretraining for robust embodied pointgoal navigation. Sensors, 2023, 23: 3553

235 Yuan L H, Xie B H, Li S. Robust test-time adaptation in dynamic scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15922–15932

236 Zhang J, Qi L, Shi Y H, et al. DomainAdaptor: a novel approach to test-time adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 18971–18981

237 Gao J Y, Yao X, Xu C S. Test-time adaptive vision-and-language navigation. In: Proceedings of International Conference on Machine Learning, 2024

238 Liu A, Huang T R, Liu X L, et al. Spatiotemporal attacks for embodied agents. In: Proceedings of the 16th European Conference on Computer Vision, 2020. 23–28

239 Ying C, Qiaoben Y, Zhou X, et al. Consistent attack: universal adversarial perturbation on embodied vision navigation. Pattern Recogn Lett, 2023, 168: 57–63

240 Tian Y P, Xu C L. Can audio-visual integration strengthen robustness under multimodal attacks? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 5601–5611

241 Chib P S, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: a survey. IEEE Trans Intell Veh, 2024, 9: 103–118

242 Huang Z Y, Liu H C, Lv C. GameFormer: game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3903–3913

243 Hu Y H, Yang J Z, Chen L, et al. Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 17853–17862

244 Shao H, Hu Y X, Wang L T, et al. LMDrive: closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 15120–15130

245 Jia X S, Gao Y L, Chen L, et al. DriveAdapter: breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 7953–7963

246 Chowdhury J, Shivaraman V, Sundaram S, et al. Graph-based prediction and planning policy network (GP3Net) for scalable self-driving in dynamic environments using deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 11606–11614

247 Ahn M, Brohan A, Brown N, et al. Do as I can, not as I say: grounding language in robotic affordances. 2022. ArXiv:2204.01691

248 Huang W L, Xia F, Xiao T, et al. Inner Monologue: embodied reasoning through planning with language models. 2022.

ArXiv:2207.05608
249 Liang J, Huang W L, Xia F, et al. Code as policies: language model programs for embodied control. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2023. 9493–9500
250 Das A, Datta S, Gkioxari G, et al. Embodied question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1–10
251 Yu L C, Chen X L, Gkioxari G, et al. Multi-target embodied question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 6309–6318
252 Driess D, Xia F, Sajjadi M S, et al. PaLM-E: an embodied multimodal language model. 2023. ArXiv:2303.03378
253 Li G, Chen X, Zhou F, et al. Self-powered soft robot in the Mariana Trench. Nature, 2021, 591: 66–71
254 Li G, Wong T W, Shih B, et al. Bioinspired soft robots for deep-sea exploration. Nat Commun, 2023, 14: 7097
255 Liu Z, Zhan W, Liu X, et al. A wireless controlled robotic insect with ultrafast untethered running speeds. Nat Commun, 2024, 15: 3815
256 Savage N. Insects offer inspiration for robot advances. Nature, 2022, 610: S18
257 Yu W, Yang C, McGreavy C, et al. Identifying important sensory feedback for learning locomotion skills. Nat Mach Intell, 2023, 5: 919–932
258 Baines R, Patiballa S K, Booth J, et al. Multi-environment robotic transitions through adaptive morphogenesis. Nature, 2022, 610: 283–289
259 Sun J, Lerner E, Tighe B, et al. Embedded shape morphing for morphologically adaptive robots. Nat Commun, 2023, 14: 6023
260 Wang D, Zhao B, Li X, et al. Dexterous electrical-driven soft robots with reconfigurable chiral-lattice foot design. Nat Commun, 2023, 14: 5067
261 Tang C, Du B, Jiang S, et al. A pipeline inspection robot for navigating tubular environments in the sub-centimeter scale. Sci Robot, 2022, 7: eabm8597
262 Palagi S, Fischer P. Bioinspired microrobots. Nat Rev Mater, 2018, 3: 113–124
263 Vaghasiya J V, Mayorga-Martinez C C, Matějková S, et al. Pick up and dispose of pollutants from water via temperature-responsive micellar copolymers on magnetite nanorobots. Nat Commun, 2022, 13: 1026
264 Urso M, Ussia M, Novotný F, et al. Trapping and detecting nanoplastics by MXene-derived oxide microrobots. Nat Commun, 2022, 13: 3573
265 Urso M, Ussia M, Pumera M. Smart micro- and nanorobots for water purification. Nat Rev Bioeng, 2023, 1: 236–251
266 Duan J, Yu S, Tan H L, et al. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Top Comput Intell, 2022, 6: 230–244
267 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359
268 Bharadhwaj H, Wang Z H, Bengio Y, et al. A data-efficient framework for training and sim-to-real transfer of navigation policies. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2019. 782–788
269 Wu J, Zhou Y, Yang H, et al. Human-guided reinforcement learning with sim-to-real transfer for autonomous navigation. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 14745–14759
270 Kadian A, Truong J, Gokaslan A, et al. Sim2Real predictivity: does evaluation in simulation predict real-world performance? IEEE Robot Autom Lett, 2020, 5: 6670–6677
271 Morad S D, Mecca R, Poudel R P K, et al. Embodied visual navigation with automatic curriculum learning in real environments. IEEE Robot Autom Lett, 2021, 6: 683–690
272 Bigazzi R, Landi F, Cornia M, et al. Out of the box: embodied navigation in the real world. In: Proceedings of International Conference on Computer Analysis of Images and Patterns, 2021. 28–30
273 Anderson P, Shrivastava A, Truong J, et al. Sim-to-real transfer for vision-and-language navigation. In: Proceedings of Conference on Robot Learning, 2021. 671–681
274 Truong J, Chernova S, Batra D. Bi-directional domain adaptation for Sim2Real transfer of embodied navigation agents. IEEE Robot Autom Lett, 2021, 6: 2634–2641
275 Savva M, Kadian A, Maksymets O, et al. Habitat: a platform for embodied AI research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 9339–9347
276 Xia F, Shen W B, Li C, et al. Interactive Gibson benchmark: a benchmark for interactive navigation in cluttered environments. IEEE Robot Autom Lett, 2020, 5: 713–720
277 Deitke M, Han W, Herrasti A, et al. RoboTHOR: an open simulation-to-real embodied AI platform. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 3164–3174
278 Rosano M, Furnari A, Gulino L, et al. On embodied visual navigation in real environments through habitat. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR), 2021. 9740–9747
279 Zhao Q, Zhang L, Wu L, et al. A real 3D embodied dataset for robotic active visual learning. IEEE Robot Autom Lett, 2022, 7: 6646–6652
280 Li C S, Zhang R H, Wong J, et al. BEHAVIOR-1K: a benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In: Proceedings of Conference on Robot Learning, 2023. 80–93
281 Cai S J, Ram A, Gou Z T, et al. Navigating real-world challenges: a quadruped robot guiding system for visually impaired people in diverse environments. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024. 1–18
282 Liang G Q, Luo H B, Li M, et al. FreeBOT: a freeform modular self-reconfigurable robot with arbitrary connection point-design and implementation. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 6506–6513
283 Dai Y, Jia L, Wang L, et al. Magnetically actuated cell-robot system: precise control, manipulation, and multimode conversion. Small, 2022, 18: 2105414
284 Dorigo M, Theraulaz G, Trianni V. Swarm robotics: past, present, and future [point of view]. Proc IEEE, 2021, 109: 1152–1165
285 Wasserman J, Yadav K, Chowdhary G, et al. Last-mile embodied visual navigation. In: Proceedings of Conference on Robot Learning, 2023. 666–678
286 Dorbala V S, Mullen J F, Manocha D. Can an embodied agent find your "cat-shaped mug"? LLM-based zero-shot object navigation. IEEE Robot Autom Lett, 2024, 9: 4083–4090
287 Kolve E, Mottaghi R, Han W, et al. AI2-THOR: an interactive 3D environment for visual AI. 2017. ArXiv:1712.05474
288 Gan C, Schwartz J, Alter S, et al. ThreeDWorld: a platform for interactive multi-modal physical simulation. 2020. ArXiv:2007.04954
289 Shen B, Xia F, Li C S, et al. iGibson 1.0: a simulation environment for interactive tasks in large realistic scenes. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021. 7520–7527
290 Orr J, Dutta A. Multi-agent deep reinforcement learning for multi-robot applications: a survey. Sensors, 2023, 23: 3625
291 Han R H, Chen S D, Hao Q. Cooperative multi-robot navigation in dynamic environment with deep reinforcement learning. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2020. 448–454
292 Marchesini E, Farinelli A. Enhancing deep reinforcement learning approaches for multi-robot navigation via single-robot

evolutionary policy search. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2022. 5525–5531
293 Chang L, Shan L, Zhang W, et al. Hierarchical multi-robot navigation and formation in unknown environments via deep reinforcement learning and distributed optimization. Robot Comput-Integrated Manuf, 2023, 83: 102570
294 Li Q B, Gama F, Ribeiro A, et al. Graph neural networks for decentralized multi-robot path planning. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 11785–11792
295 Tolstaya E, Paulos J, Kumar V, et al. Multi-robot coverage and exploration using spatial graph neural networks. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021. 8944–8950
296 Blumenkamp J, Morad S, Gielis J, et al. A framework for real-world multi-robot systems running decentralized GNN-based policies. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2022. 8772–8778
297 Collins L, Ghassemi P, Esfahani E T, et al. Scalable coverage path planning of multi-robot teams for monitoring non-convex areas. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2021. 7393–7399
298 Liu X, Guo D, Liu H, et al. Multi-agent embodied visual semantic navigation with scene prior knowledge. IEEE Robot Autom Lett, 2022, 7: 3154–3161
299 Okumura K. LaCAM: search-based algorithm for quick multi-agent pathfinding. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 11655–11662
300 Gao Y M, Wang Y J, Zhong X G, et al. Meeting-merging-mission: a multi-robot coordinate framework for large-scale communication-limited exploration. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. 13700–13707
301 Liu X, Guo D, Zhang X, et al. Heterogeneous embodied multi-agent collaboration. IEEE Robot Autom Lett, 2024, 9: 5377–5384
302 Chen W, Zhou S, Pan Z, et al. Mapless collaborative navigation for a multi-robot system based on the deep reinforcement learning. Appl Sci, 2019, 9: 4198
303 Yan C, Xiang X J, Wang C, et al. Flocking and collision avoidance for a dynamic squad of fixed-wing UAVs using deep reinforcement learning. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021. 4738–4744
304 Wu L, Guo B, Xu R N, et al. Emergence of crowd modular robotics: a ubiquitous computing perspective (in Chinese). Sci Sin Inform, 2023, 53: 2107–2151
305 Duan H, Huo M, Fan Y. From animal collective behaviors to swarm robotic cooperation. Natl Sci Rev, 2023, 10: nwad040
306 Liu S, Guo B, Fang C, et al. Enabling resource-efficient AIoT system with cross-level optimization: a survey. IEEE Commun Surv Tut, 2024, 26: 389–427
307 Wehner R. Desert ant navigation: how miniature brains solve complex tasks. J Comp Physiol A-Sensy Neural Behav Physiol, 2003, 189: 579–588
308 Webb B, Wystrach A. Neural mechanisms of insect navigation. Curr Opin Insect Sci, 2016, 15: 27–39
309 Raibert M, Blankespoor K, Nelson G, et al. BigDog, the rough-terrain quadruped robot. IFAC Proc Volumes, 2008, 41: 10822–10825
310 Jeong K H, Kim J, Lee L P. Biologically inspired artificial compound eyes. Science, 2006, 312: 557–561
311 Huang H, He W, Wang J, et al. An all servo-driven bird-like flapping-wing aerial robot capable of autonomous flight. IEEE ASME Trans Mechatron, 2022, 27: 5484–5494
312 de Croon G C H E, Dupeyroux J J G, Fuller S B, et al. Insect-inspired AI for autonomous robots. Sci Robot, 2022, 7: eabl6334
313 Borst A. Drosophila's view on insect vision. Curr Biol, 2009, 19: R36–R47
314 Wu T Y. Fish swimming and bird/insect flight. Annu Rev Fluid Mech, 2011, 43: 25–58
315 Seth A, Hicks J L, Uchida T K, et al. OpenSim: simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. PLoS Comput Biol, 2018, 14: e1006223
316 Sandamirskaya Y, Kaboli M, Conradt J, et al. Neuromorphic computing hardware and neural architectures for robotics. Sci Robot, 2022, 7: eabl8419
317 Yamazaki K, Vo-Ho V K, Bulsara D, et al. Spiking neural networks and their applications: a review. Brain Sci, 2022, 12: 863
318 Oruma S O, Sánchez-Gordón M, Colomo-Palacios R, et al. A systematic review on social robots in public spaces: threat landscape and attack surface. Computers, 2022, 11: 181
319 Guo R X, Tian J W, Wang B H, et al. Cyber-physical attack threats analysis for UAVs from CPS perspective. In: Proceedings of International Conference on Computer Engineering and Application (ICCEA), 2020. 259–263
320 Bezzo N, Weimer J, Pajic M, et al. Attack resilient state estimation for autonomous robotic systems. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014. 3692–3698
321 Deng G L, Zhou Y, Xu Y, et al. An investigation of Byzantine threats in multi-robot systems. In: Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses, 2021. 17–32
322 Wan Z W, Shen J J, Chuang J, et al. Too afraid to drive: systematic discovery of semantic dos vulnerability in autonomous driving planning under physical-world attacks. 2022. ArXiv:2201.04610
323 Cancelli E, Campari T, Serafini L, et al. Exploiting proximity-aware tasks for embodied social navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 10957–10967
324 Vitale J, Tonkin M, Herse S, et al. Be more transparent and users will like you: a robot privacy and user experience design experiment. In: Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, 2018. 379–387
325 Gu J, Stefani E, Wu Q, et al. Vision-and-language navigation: a survey of tasks, methods, and future directions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 7606–7623
326 Vasylkovskyi V, Guerreiro S, Sequeira J. BlockRobot: increasing privacy in human robot interaction by using blockchain. In: Proceedings of IEEE International Conference on Blockchain, 2020. 106–115
327 Chatzimichali A, Harrison R, Chrysostomou D. Toward privacy-sensitive human-robot interaction: privacy terms and human-data interaction in the personal robot era. Paladyn J Behav Robot, 2021, 12: 160–174
328 Zhou K W, Wang X E. FedVLN: privacy-preserving federated vision-and-language navigation. In: Proceedings of European Conference on Computer Vision, 2022. 682–699