Computers & Graphics 130 (2025) 104256

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

3D Reconstruction in Robotics: A Comprehensive Review☆

Dharmendra Selvaratnam*, Dena Bazazian

University of Plymouth, Faculty of Science and Engineering, School of Engineering, Computing and Mathematics (SECaM), United Kingdom

ARTICLE INFO

Keywords:
3D reconstruction
Robotics
Computer vision
Computer graphics

ABSTRACT

In this paper, we delve into the swiftly progressing domain of 3D reconstruction within robotics, a field of critical importance in contemporary research. The inherent potential of robots to engage with and understand their environment is significantly enhanced by integrating 3D reconstruction techniques, which draw inspiration from the complex processes of natural evolution and human perception. This study not only highlights the importance of 3D reconstruction methodologies in the broader context of technological advancement but also outlines their pivotal contributions to the field of robotics. Humans have evolved over millions of years to adapt to their surroundings through natural selection, enabling them to perceive the world. 3D reconstruction methods are inspired by natural processes to replicate objects, providing more detailed information about the perceived object. With this approach to object perception, robotics plays a crucial role in utilising these techniques to interact with the real world. Our study illustrates recent advancements in applying 3D reconstruction methods within robotics and discusses necessary improvements and applications for future research in the field.

1. Introduction

3D reconstruction significantly advances robotics by bridging simple perception and sophisticated interaction with the environment. This technology equips robots with a complete understanding of their surroundings, facilitating critical tasks such as navigation [1], manipulation [2], and autonomous decision making [3]. It enables robots to accurately measure distances [4], recognise objects [5], and plan safe and effective actions in complex and dynamic environments.

Neural Radiance Fields (NeRFs) have become a formidable method for capturing precise scene geometries, offering the ability to synthesise novel views of intricate scenes with remarkable detail. This technique, along with advances in Simultaneous Location and Mapping (SLAM) [6], which provides real-time environment mapping and localisation, has greatly improved the autonomy of robotic systems. Such state-of-the-art developments in 3D reconstruction, including work by researchers such as [7–10], are pivotal for robotic applications.

Our work thoroughly analyses these advanced 3D reconstruction techniques and their impact on robotic perception and action. We focus on the synergistic integration of NeRF, SLAM [11], and other innovative methods, detailing how they collectively improve robotic capabilities. We highlight the importance of accurate 3D reconstructions, such as NeRF [12] and SLAM, which allow robots to perceive and interact better with their environment [13]. Finally, we also add some works that used Gaussian-Splatting [14] which is gaining popularity with NeRF models due to its high speed.

Furthermore, our paper discusses the utility of different 3D modelling techniques such as surfels and polygonal modelling. Surfels, which use oriented discs to represent 3D surfaces, are excellent for quickly rendering complex, detailed surfaces, and are beneficial in real-time rendering and point-cloud processing. Conversely, polygonal modelling, which employs vertices and edges to form polygons, is superior in applications requiring precise geometric manipulations, like computer-aided design (CAD) and animation, with the choice between these methods depending on the specific needs such as speed or geometric detail. To our knowledge, there is no other comprehensive review paper that focuses on 3D reconstruction techniques in robotics.

We structure our paper to guide the reader through the intricate landscape of 3D reconstruction in robotics. It begins with an overview of the fundamental principles, explores cutting-edge innovations, and examines case studies where advanced reconstruction methods have transformed robotic applications. The paper concludes with a discussion of potential future developments, the broader impact of our findings, and an exploration of ongoing challenges and research opportunities, paving the way for future advancements in robotic 3D reconstruction.

☆ This article was recommended for publication by Yulan Guo.

* Corresponding author.
E-mail address: dharmendra.selvaratnam@plymouth.ac.uk (D. Selvaratnam).

https://doi.org/10.1016/j.cag.2025.104256
Received 26 January 2025; Received in revised form 16 April 2025; Accepted 13 May 2025
Available online 30 May 2025

2. Related work

The field of 3D reconstruction has seen significant advancements, with numerous review and survey papers providing insight into various techniques and their applications across diverse domains such as archaeology, medicine, virtual reality, and robotics. This section synthesises key literature, highlighting both broad surveys and focused reviews that contribute to our understanding of 3D reconstruction technologies.

2.1. Comprehensive surveys on 3D reconstruction

Several comprehensive surveys have been published covering a wide range of 3D reconstruction methods. For example, Picard et al. [15] provide a broad overview of various 3D reconstruction technologies, discussing their applications across multiple fields. Similarly, Samavati et al. [16] focus on deep learning approaches in 3D reconstruction, categorising methods based on input modalities and their effectiveness. Macario et al. [17] also offer a comprehensive examination of state-of-the-art techniques, emphasising the evolution of these methods over time. Collectively, these works illustrate the breadth of methodologies available in 3D reconstruction and their relevance to contemporary challenges.

2.2. Focused reviews in specific areas

In addition to broad surveys, there are narrower reviews that go into specific frameworks or technologies. For example, Cai et al. [18] present an in-depth analysis of the SLAM (Simultaneous Localisation and Mapping) framework, detailing its core modules and the challenges faced by SLAM networks. This review is particularly useful for understanding the intricacies involved in robot-based 3D reconstruction. Similarly, Zhang et al. [19] explore the challenges and applications of 3D LED technology within SLAM systems, providing information on recent advances and ongoing issues. These focused reviews are beneficial for practitioners seeking detailed knowledge about particular aspects of 3D reconstruction.

2.3. Our approach

Our survey distinguishes itself by concentrating on the integration of advanced technologies such as Neural Radiation Fields (NeRF) and SLAM within robotic systems. By focusing on these cutting-edge approaches, we aim to enhance robotic perception and interactivity. This targeted perspective allows us to explore how these technologies synergistically improve robotic capabilities without sacrificing depth or breadth. Our review builds upon existing methodologies by emphasising their applicative enhancements in robotics. Pioneering works like those by Saeedi et al. [20] and Maboudi et al. [21] have laid the groundwork for understanding SLAM and stereo reconstruction methods, while recent advancements extend this discussion to include dense scene reconstructions using multiview and monocular approaches. By integrating these methodologies, our work presents a comprehensive view of how various 3D reconstruction techniques complement each other to enhance robotic perception in dynamic environments.

Recent developments highlight the importance of real-time data processing in robotics. Liu et al. [22] demonstrate significant advances in autonomous driving through edge computing, which aligns with the sensory requirements outlined by Zaffar et al. [23] for effective implementation of SLAM. The integration of LiDAR and RGB-D cameras, as discussed by Kang et al. [24] and Cui et al. [25], provides crucial insights into indoor and outdoor scene reconstructions. This fusion not only enhances object recognition but also improves scene understanding, which is vital for real-time robotic applications.

In synthesising these works, our paper not only discusses individual technological advancements but also highlights the synergies between different approaches. By focusing on the integration of NeRF and SLAM within robotics, we provide a comprehensive perspective on the state-of-the-art in robotic 3D reconstruction. This integrated approach is essential for the development of robots capable of navigating complex environments while adapting intelligently to spatial dynamics. The reviewed literature underscores the transformative potential of advanced 3D reconstruction techniques in robotics, paving the way for future research directions that could further enhance robotic capabilities through sophisticated modelling techniques (see Table 1).

Table 1
Summary table of SLAM main papers used in this section.

Paper	Sensor	Year	Platform
ORB-SLAM2 [26]	Stereo Camera	2017	CPU
ORB-SLAM3 [27]	Monocular Camera	2020	CPU
OpenVSLAM [28]	Stereo Camera	2019	CPU
RTABMap [29]	Stereo Camera	2015	CPU
Air-SLAM [30]	Monocular Camera	2023	CPU
UV-SLAM [31]	Monocular Camera	2023	CPU
iSLAM [32]	Monocular Camera	2023	CPU
SLAM-Box [33]	LiDAR	2021	FPGA + CPU
RGBD-SLAM-Agriculture [34]	RGB-D Camera	2022	CPU
UW-Vis-SLAM [35]	Visual & Acoustic (UUV)	2023	CPU

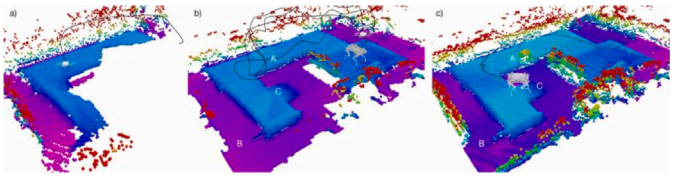


Fig. 1. Three snapshots of the estimated elevation map are shown [42]. In the first image, the map is incrementally constructed using dense depth estimates from the flying robot's camera. The second image displays the map after the flying robot has observed the entire scene, which is then used for an initial navigation plan for the legged robot. The final image shows the map after fusing data from the laser range sensor on the walking robot with the original map. Point A indicates a newly added obstacle after the flying robot's initial mapping.

3. Simultaneous Localisation and Mapping (SLAM)

3.1. Overview

Simultaneous Location and Mapping (SLAM) allows devices to create a 3D map while localising themselves within it and calculating trajectories for autonomous navigation. This technique is crucial in robotics for generating sparse 3D point cloud maps in real-time, enabling robots to adapt to their surroundings.

SLAM has a wide range of applications, such as in underwater [35–37], planetary [33], and indoor environments [33,38]. In addition, it can use different sensors to create its environment, such as RGB-D [39,40] and LiDAR [33,39]. When considering storage, it can process data using CPU storage [33] or in the cloud [40,41]. Finally, as shown in Fig. 1, multiple robots can be used to plan the trajectory in a unified environment to improve the quality of the trajectory [40].

We have provided in-depth applications and techniques of SLAM described above in the remaining sections.

3.2. Indoor environment reconstructions

The integration of SLAM techniques in various applications demonstrates significant advancements in autonomous navigation and environment reconstruction. For example, SLAM is used to build drone trajectories within a 3D point cloud, improving trajectory accuracy by filtering out outlier data, although it is sensitive to light fluctuations that affect overall positioning [38]. Building on this, a combination of FPGA and CPU-based computing generates SLAM reconstructions of

indoor environments using LiDAR to create point-cloud maps and TSDF Fusion for dense reconstructions. This approach yields high-quality reconstructions but requires expensive and custom hardware [33]. Similarly, edge computing is utilised for real-time scene reconstruction, processing data on a powerful edge device, and fusing maps in the cloud. This method mirrors the peer-to-peer approach but struggles with cluttered scenes, indicating the need for improved clutter handling [41]. Other research focuses on large-scale applications of SLAM in complex environments such as factory halls or construction sites. It uses a robot system equipped with four stereo cameras and a 3D laser scanner to collect data in a large factory floor, comparing the LiDAR and visual SLAM approaches [43]. Specifically, when we consider indoor reconstructions, dynamic elements will always cause problems as they will move independently of the observer, to remove these objects and reconstruct based on static objects [44] proposed a method to eliminate dynamic elements by exploiting geometric residuals.

3.3. SLAM in underwater

SLAM is also extended to underwater environments using unmanned underwater vehicles (UUV), combining ORB-SLAM2 (Oriented FAST and Rotated BRIEF-SLAM) with an acoustic odometer from a Doppler Velocity Log (DVL) gyroscope to allow sparse reconstructions of submerged surfaces, with potential improvements through dense reconstructions [35]. Recently, RU-SLAM [36] introduced a system to address underwater SLAM challenges. It features UWNet (Under Water Network), a specialised generator that improves keypoint feature extraction in weak textures and degraded images. UWNet is integrated into ORB-SLAM3, enhancing accuracy and robustness in complex underwater scenarios. The integration of acoustic and visual sensors improves the performance of SLAM in underwater environments, addressing challenges such as low visibility and weak textures. Depth mapping combined with semantic understanding improves autonomous localisation when GPS is unavailable. In complex underwater environments, the integration of DVL, gyroscopes, and altimeters significantly improves SLAM for autonomous underwater vehicles (AUVs) [37]. This multisensor approach improves navigation and mapping accuracy by combining acoustic and optical data, ensuring precise localisation and robust environmental mapping. The methodology demonstrates how sensor fusion can address underwater navigation challenges, providing reliable performance in intricate underwater settings. Recently, [36] introduced RU-SLAM, a system designed to address underwater SLAM challenges. It features UWNet (Under Water Network), a specialised generator that improves keypoint feature extraction in weak textures and degraded images. UWNet is integrated into ORB-SLAM3, enhancing accuracy and robustness in complex underwater scenarios. A multisensor approach for underwater SLAM integrates acoustic sensors (DVL, gyroscopes, altimeters) with visual data, significantly enhancing the navigation and mapping accuracy of AUVs in complex underwater environments by combining acoustic and optical data [45]. Furthermore, [46] proposes a novel underwater SLAM system using a light field camera and an inertial measurement unit (IMU). The system addresses challenges specific to underwater environments, such as refraction and low visibility, by leveraging the unique capabilities of light-field imaging.

3.4. Constrained resource SLAM

In low-memory devices, a SLAM system capable of sparse reconstructions on robots produces dense voxel-based maps from sparse SLAM data, demonstrating the efficiency for large-scale reconstructions despite some accuracy limitations [47]. A multirobot system for scene reconstruction using SLAM distributes computations to edge devices and merges maps in the cloud. This system, tested in simulations and real-world experiments, faces challenges in map accuracy due to incomplete scans [40]. Using advanced SLAM applications, a small robotic rover with a stereo camera generates accurate 2D maps of disaster scenes. This approach employs standard stereo cameras for quick replacements, although a monocular setup and 3D SLAM could provide more detailed scene information [13]. Expanding on these methodologies, integrating LiDAR and RGB-D sensors achieves 3D dense reconstruction using a small tracked robot. This hybrid approach constructs a 2D SLAM map with LiDAR and then overlays RGB-D data to refine the reconstruction, although computational demands present a challenge for real-time applications [39]. SLAM is applied to UUVs for vessel hull inspections using monocular cameras to create dense piecewise-planar SLAM reconstructions. This method, effective for depth analysis, could be enhanced by point cloud reconstructions to detect hull defects more reliably [48].

Collaborative SLAM techniques that use two robots, a drone and a ground quadrupedal robot, allow collaborative trajectory planning in chaotic environments. The drone produces the initial SLAM point map, voxelised to show the environment in relation to both robots.

3.5. Collaborative SLAM techniques

Collaborative trajectory planning around a chaotic environment was achieved with a drone and a ground quadrupedal robot by Fankhauser et al. [42], as shown in Fig. 1. The drone produces the initial SLAM point map, which is voxelised to show the environment in relation to the drone and the quadrupedal robot, to build an initial trajectory for the quadrupedal robot. Then, the quadrupedal robot performs the initialised trajectory and updates the trajectory and map with the onboard RGB-D camera on the quadrupedal robot if there are unseen objects in the scene. The paper does not show the robot in an outdoor environment, nor how the algorithm copes with outdoor scenery.

3.6. Multi camera SLAM

Compared to single-camera simultaneous location and mapping (SLAM) systems, multi-camera SLAM [49] presents several significant advantages. One of the primary benefits is the enhanced field of view achieved through the use of multiple cameras, which allows the simultaneous capture of various perspectives. This expanded coverage notably mitigates the risk of losing track of landmarks, particularly in dynamic or cluttered environments.

In addition to a wider field of view, multi-camera configurations facilitate better depth estimation [50]. Stereo or multi-camera setups inherently obtain depth information via triangulation, which leads to superior three-dimensional reconstruction and greater localisation accuracy, especially for objects positioned at varying distances.

Moreover, multi-camera SLAM exhibits increased robustness to occlusion. If an object obstructs one camera's view, the remaining cameras can still track features, thereby ensuring more reliable performance within complex settings. Furthermore, multi-camera systems are adept at managing dynamic environments, as they are better equipped to differentiate between moving objects and static elements. In contrast, single-camera SLAM often encounters challenges in this regard, which can result in significant localisation errors.

While several researchers [49,51,52] have identified solutions for multi-camera SLAM, none have successfully integrated these solutions into large robotic systems for practical, real-world applications. A multi camera practical system, as introduced in [53], uses a monocular camera configuration to achieve joint localisation through a main localisation module and a sub-vocalisation module. Additionally, a depth estimation method based on camera geometry is proposed to create an initial depth map, which is later refined using a convolutional neural network.

According to [54], even the use of multiple cameras poses a significant issue of whether all cameras capture images simultaneously due to their different shutter speeds. There needs to be a system that aligns the images to a continuous time frame. By incorporating a continuous-time

motion model, the framework accurately relates information across asynchronous multi-frames during tracking, local mapping, and loop closing, enhancing the precision of trajectory estimation.

Even with a synchronous camera system, some adjustments and matching are needed to enhance localisation accuracy. The introduction of a virtual camera, termed a BundledFrame [55], allows for the seamless integration of measurements from all cameras, facilitating effective data fusion. The BundledFrame concept in BundledSLAM involves creating a virtual camera that consolidates measurements from multiple physical cameras. This approach facilitates effective data fusion by mapping all camera measurements onto a unified framework, enhancing the robustness and accuracy of feature matching and place recognition. The BundledFrame is meticulously designed to seamlessly adapt to various multi-camera configurations, allowing for efficient integration of data captured from multiple cameras. Furthermore, by incorporating extrinsic parameters in the bundle adjustment process, the system achieves precise trajectory estimation, further improving the performance of the SLAM system. The system extends the capabilities of the state-of-the-art ORB-SLAM2 by incorporating pose estimation and map reuse from multiple cameras, enhancing its adaptability to multicamera configurations. Experimental evaluations using EuRoC [56] datasets demonstrate that BundledSLAM achieves superior accuracy compared to existing approaches, confirming the effectiveness of its multi-camera integration and optimisation strategies.

Though we have multiple camera systems and can either use synchronised or asynchronous camera systems, there comes a question of how well we can place these cameras to obtain the maximum field of view to capture as much of the environment as we can. This problem was solved using non-overlapping cameras in SLAM [57]. Non-overlapping cameras capture diverse viewpoints, enriching environmental understanding. This reduces susceptibility to occlusions and feature-poor areas by combining multiple perspectives. Integrating data from various angles leads to more precise pose estimation and mapping.

Multi Camera-Slam was used for a specific application using a novel approach [58] to improve simultaneous localisation and mapping (SLAM) in challenging off-road environments. A generalised camera model is utilised to project spatial points onto a spherical imaging plane, facilitating accurate depth estimation even with non-standard camera configurations. This design enhances environmental perception, crucial for unmanned ground vehicles [59] navigating complex terrains.

3.7. SLAM in the view of NeRF and splatting

To achieve rapid rendering quality, [60] introduced Gaussian representation for the first time, contrasting with neural representation and other methodologies. This choice of volumetric representation is justified as it facilitates efficient rendering while enabling the delineation of boundaries via a silhouette map, which can be readily generated by aggregating the opacity of the Gaussian volumes present in the scene. Furthermore, this approach allows for straightforward updates to the map by adding Gaussians in previously unseen regions as one navigates from one position to another. Given that camera motion can be conceptualised as maintaining a stationary camera while the scene itself shifts, there exists a direct gradient in relation to the camera parameters, thereby enabling rapid optimisation. In a similar vein, [61] employed Gaussian plating [62], utilising a coarse-to-fine strategy to select reliable 3D Gaussians to optimise camera pose estimations. The system commences with a broad alignment and progressively refines the pose estimation by focusing on finer details. By concentrating computational resources on the most informative Gaussians, this system not only reduces processing time but also enhances the robustness of camera pose estimation, ultimately leading to more stable and accurate tracking. GS-SLAM has demonstrated competitive performance compared to existing state-of-the-art real-time methodologies, as evidenced by its results in datasets such as Replica [63] and TUM-RGBD [64]. To mitigate the "forgetting" problem — where continuous mapping can induce overfitting to recent frames, thereby degrading the quality of earlier reconstructions — the system integrates additional regularisation parameters [65]. These parameters guide the optimisation process, fostering consistency across all frames. Specifically, the employed regularisation technique incorporates a loss term that penalises deviations from previously optimised Gaussian parameters. Consequently, this regularisation ensures that updates to the map do not compromise the integrity of previously reconstructed areas, thus maintaining overall rendering quality and contributing to a more stable and reliable SLAM system. Although previous SLAM systems have utilised 3DGS for passive mapping, AG-SLAM [66] is the first to incorporate this representation into an active SLAM framework [67]. This integration allows the system to autonomously explore and map unknown environments, leveraging the strengths of 3DGS for real-time, high-quality scene reconstruction. The system formulates path planning as an active learning problem, generating multiple feasible paths and employing an uncertainty-aware algorithm to select the optimal one. This approach balances exploration (gathering new information) with localisation accuracy (minimising potential errors in position estimation).

Traditional NeRF-based SLAM systems often rely on a single multilayer perceptron (MLP) to represent the entire scene, which can lead to over-smoothed reconstructions and challenges in scaling to larger environments. NICE-SLAM [68] addresses these limitations by introducing a grid-based hierarchical representation that allows localised updates and better scalability. This structure enables more detailed and accurate reconstructions compared to methods that utilise a global neural scene encoding. The system used multi-level grid-based features to represent scene geometry and appearance. This hierarchical structure captures details at various spatial resolutions, enabling the system to model both coarse and fine aspects of the environment. Similarly, SLAIM [69] applies a Gaussian pyramid filter to the NeRF representation, creating multiple levels of detail. This facilitates a coarse-to-fine optimisation strategy during camera tracking, starting with broad alignments at lower resolutions and progressively refining at higher resolutions. This hierarchical approach addresses the narrow basin of attraction in image space optimisation, reducing susceptibility to local minima, and enhancing convergence. By mitigating the lack of initial correspondences, it improves the robustness and accuracy of camera pose estimation. It also introduces a new target-ray termination distribution for improved geometry convergence. Considering rays, RoDyn-SLAM [70] introduces a novel approach to Simultaneous Localisation and Mapping (SLAM) in dynamic environments by integrating Neural Radiance Fields (NeRF) with advanced motion mask generation and pose optimisation techniques. It addresses the challenges of dynamic scenes by generating motion masks that filter out invalid sampled rays. This is achieved by fusing optical flow masks, which detect motion between frames, with semantic masks that identify object categories. By combining optical flow and semantic information, the system accurately distinguishes between static and dynamic elements within the scene. This fusion improves motion detection precision, allowing the SLAM system to focus on stable features for mapping and localisation, thus improving robustness in dynamic settings. Another approach called DNS SLAM also exploited semantic information by using a hybrid representation that combines implicit neural fields with explicit semantic segmentation [71]. By integrating semantic information directly into the scene representation, DNS SLAM can produce class-wise decomposed reconstructions. This capability is particularly beneficial for applications requiring detailed scene understanding, such as robotics and augmented reality.

For example, as real world applications, Vial et al. [72] developed a two-dimensional pose SLAM system for Autonomous Underwater Vehicles (AUVs) that utilises a dead reckoning system based on Lie Theory and a rigid scan matching technique tailored for acoustic data. Their system was tested with real data, demonstrating successful real-time execution and effective mapping of underwater environments.

Table 2
Quantitative results of various SLAM approaches, categorised into onboard, edge-based, and cloud-based methods. Metrics include collected data size, runtime, power consumption, and refresh rate. The table compares different SLAM implementations based on these metrics to highlight performance differences and trade-offs.

Papers	Coll. data size (MB)	Runtime (s)	Power (W)	Refresh rate (Hz)
Onboard SLAM Approaches				
SLAM-Box [33]	1.52494 (Mean)	0.038	13.8	10
Under Water SLAM [35]	–	–	–	5.3 and 20.8 (Mean)
Topomap [47]	4.320 (Max)	86.1 (Max)	–	20–30
Edge-based SLAM Approaches				
Collaborative Navigation [42]	–	~ 0.02	–	5
DOOR-SLAM [38]	1.56189	–	–	6
Cloud-based SLAM Approaches				
Offloading SLAM [41]	–	–	6	10
RecSLAM [40]	–	~ 0.039 (10)	–	5.5–10

Furthermore, the work by Mar. Sci. Eng. [73] introduces a novel registration algorithm that enhances the accuracy of feature matching in sparse underwater environments by integrating the Inertial Measurement Unit (IMU) and Doppler velocimeter data for global state estimation. These examples underscore the practical applicability and success of SLAM methodologies in underwater robotics, showcasing their ability to navigate and map complex aquatic landscapes effectively.

3.8. Quantitative results

The experimental results from the SLAM papers analysed in this section [33,35,38,40–42,47] are shown in Table 2. However, some papers [13,39,48] have not provided much experimental data in their papers, making the papers not quantifiable.

The use of cloud computing in SLAM facilitates further distribution of the processing workload. For example, in [40,41], the LiDAR Sensor RPLiDAR A1 is used to capture the local environment due to its small range and cost-effectiveness. These papers [40,41] detail how data from multiple robots are processed on an edge computer, with [40] noting that adding more robots to the network increases the computational load and computation time, as shown in Table 2. Furthermore, [41] reports an average power consumption of 4.1 W when the system performs various tasks.

In SLAM, custom wide-view systems are often used to enhance environmental reconstruction. For example, the ASL VI-Sensor [74] has been used in multiple SLAM studies [42,47].

It was implemented on a quadrupedal robot and a monocular camera on a drone to collaboratively perform SLAM, incorporating perspectives from each robot [42]. Alternatively, it was used on a single robot to produce sparse SLAM reconstructions, taking around 20 ms per scan in one setup, while another method required 86.1 s for complex environment reconstructions, with the largest data size being 4.3 MB [47].

High-quality and long-range scanning capabilities of Velodyne VLP-16 LiDAR sensors have been highlighted in works [33,38]. The high refresh rates and extended range of these sensors enable precise reconstructions and trajectory predictions. The Intel RealSense D435 was used in aerial applications for trajectory prediction, while the LiDAR sensor facilitated the navigation of multiple robots in subterranean environments, demonstrating storage efficiency with 1.5MB used and a mean Average Translation Error (ATE) of 9.5857 [38]. Another application used the LiDAR sensor for dense scene reconstructions, distributing the computational load between an FPGA and a CPU for TSDF reconstruction, achieving an average scan time of 38 ms with a higher power consumption of 13.8 W, resulting in 1.5 MB of data per scan [33].

A novel approach combining visual and acoustic odometry for underwater reconstructions was presented [35]. Experiments with two different Unmanned Underwater Vehicles (UUVs) showcased varying mean refresh rates and highlighted a mean Root Mean Squared Error (RMSE) of 0.27 on the absolute trajectory of the submersible robots. High speeds and lighting conditions were found to cause failures in the ORB-SLAM2 algorithm, underscoring the challenges of maintaining accurate reconstructions under these conditions (see Table 3).

Table 3
SLAM Papers, Datasets, and Metrics: All values represent averages over sequences. Metrics include: TRMSE (Translation RMSE), ATE (Absolute Trajectory Error), ATE RMSE (Root Mean Square of ATE), and KLE (Keyframe Localisation Error).

Dataset	Metric	Paper	Camera type	Values
KITTI	TRMSE	ORB-SLAM2 [26]	Stereo	2.81
		ORB-SLAM3 [27]	Monocular	16.878
		OpenVSLAM [28]	Stereo	17.932
		RTABMap [29]	Stereo	2.817
	ATE	OpenVSLAM [28]	Stereo	56.32
		ORB-SLAM [75]	Stereo	68.78
EuRoC	TRMSE	ORB-SLAM2 [26]	Stereo	0.0435 (V2 03 failed)
		Air-SLAM [30]	Monocular	0.030
		iSLAM [32]	Monocular	0.508 (loop)
		UV-SLAM [31]	Monocular	0.139 (loop)
		Kimera [76]	Monocular	0.114 (loop)
		PL-SLAM [77]	Stereo	0.061 (loop)
	ATE RMSE	DSO [78]	Monocular	0.601
		SVO [79]	Monocular	0.294
		DSM [80]	Monocular	0.126
		ORB-SLAM3 [27]	Monocular	0.041
	ATE	OpenVSLAM [28]	Monocular	23.84
		ORB-SLAM [75]	Monocular	27.96
TUM RGB-D	TRMSE (fr1)	ORB-SLAM2 [26]	Monocular	0.018
		DPV-SLAM [81]	Monocular	0.076
		DeepTAM [82]	Monocular	0.116
		TartanVO [83]	Monocular	0.206
		DeepV2D [84]	Monocular	0.375
		DeepFactors [85]	Monocular	0.233
	ATE RMSE	Photo-SLAM [86]	RGB-D	1.3
		NICE-SLAM [68]	RGB-D	4.0
		Point-SLAM [87]	RGB-D	2.6
		SplaTAM [60]	RGB-D	3.2
		RGBD GS-ICP [88]	RGB-D	2.4
	KLE	ORB-SLAM [75]	Monocular	1.47
		PTAM [89]	Monocular	0.675 (failed)
		LSD-SLAM [90]	Monocular	14.96
		RGBD SLAM [91]	Monocular	2.51

A quantitative comparison of SLAM performance across different datasets, metrics, and papers is provided, and for the KITTI dataset, stereo camera setups (e.g., ORB-SLAM2 and RTABMap) achieve significantly lower TRMSE values (around 2.81 and 2.817) compared to monocular approaches (e.g., ORB-SLAM3 with a TRMSE of 16.878), highlighting the advantage of stereo data for trajectory estimation, as shown in Table 3. Absolute Trajectory Error (ATE) also shows better results for stereo methods, such as RTABMap (2.817), compared to monocular methods like OpenVSLAM (17.932).

For the EuRoC dataset, monocular SLAM methods like Air-SLAM and UV-SLAM achieve impressively low T-RMSE values (0.03 and 0.139, respectively), indicating strong performance in tightly controlled indoor environments. However, some monocular methods, such as iSLAM (0.508), show larger errors, possibly due to difficulty handling specific sequences. Stereo methods like ORB-SLAM2 demonstrate comparable performance with very low T-RMSE values (0.0435). ATE values for monocular approaches vary significantly, with OpenVSLAM and ORB-SLAM showing higher errors (23.84 and 27.96, respectively), suggesting challenges in robust trajectory estimation for monocular setups.

For the TUM-RGBD dataset, RGB-D methods outperform monocular approaches. For instance, NICE-SLAM achieves an ATE RMSE of 2.6, significantly better than monocular methods like DeepV2D (0.375). Similarly, KLE values are better for RGB-D methods like RGBD GS-ICP SLAM (2.4) compared to monocular methods like ORB-SLAM (1.47). This highlights the importance of depth data in improving trajectory and localisation accuracy.

Overall, the data shows a clear trend where stereo and RGB-D systems outperform monocular approaches across all datasets in metrics like TRMSE, ATE, and KLE, underscoring the value of richer sensor modalities for SLAM tasks. Monocular methods, while more lightweight and flexible, demonstrate a broader range of performance, indicating sensitivity to dataset characteristics and implementation differences (see Table 3).

4. Gaussian splatting in robotics

4.1. Overview

Gaussian Splatting is a technique used for 3D reconstruction, rendering, and data representation. It represents surfaces or volumes using Gaussian kernels instead of traditional meshes, point clouds, or voxel grids [92]. This approach is highly efficient for processing sparse or dense data and enables smooth representation and rendering of complex 3D structures. Gaussian splatting outperforms traditional methods like Neural Radiance Fields (NeRF), voxel grids, and mesh-based Signed Distance Functions (SDFs) in terms of speed due to its representation efficiency, computational simplicity, and scalable rendering pipeline. The real-time rendering capability of Gaussian splatting facilitates the rapid reconstruction of environmental contexts by robots, allowing for swift adaptation to dynamic changes in their surroundings in the environment quickly.

4.2. Theory

Gaussian splatting is a novel way of representing a 3D scene that models the scene as a continuous density field composed of Gaussian primitives. These primitives are parameterised by their mean position ($\mu \in \mathbb{R}^3$), covariance matrix ($\Sigma \in \mathbb{R}^{3\times3}$), and amplitude ($A \in \mathbb{R}$). The scene is represented as a summation of these Gaussian components, with the density at any spatial point $\mathbf{x} \in \mathbb{R}^3$ defined as:

$$S(\mathbf{x}) = \sum_{i=1}^{N} A_i \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right),$$

where N is the total number of Gaussians, Σ_i^{-1} is the inverse covariance matrix of the ith Gaussian, and μ_i represents the mean position or centre of the Gaussian. Each Gaussian contributes smoothly to the overall density field, allowing the method to avoid grid-like artifacts common in discretised representations.

For rendering, Gaussian splatting projects these 3D Gaussians onto a 2D image plane. The projected density $\mathcal{P}(\mathbf{u})$ at a pixel position $\mathbf{u} \in \mathbb{R}^2$ on the image plane is computed by integrating the density of all Gaussians along the viewing ray:

$$\mathcal{P}(\mathbf{u}) = \int \sum_{i=1}^{N} A_i \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right) d\mathbf{x}.$$

Table 4
Summary of main Gaussian Splatting-related robotics papers used in this section.

Paper	Sensors used	Year	Platform
Street Gaussians	Cameras	2023	GPU
3DGS	Cameras	2023	GPU
HUGS	Cameras	2024	GPU
HF-SLAM	RGB-D	2024	GPU
E-SLAM	RGB-D	2023	GPU
Point-SLAM	RGB-D	2023	GPU
DDN-SLAM	RGB-D	2024	GPU
NeSLAM	RGB-D	2024	GPU
Gaussian-SLAM	RGB-D	2023	GPU
SplaTAM	RGB-D	2024	GPU
ManiGaussian	RGB, Proprioception (RLBench)	2025	GPU
GNFactor	RGB, Proprioception (RLBench)	2023	GPU
LSeg	RGB	2022	GPU
Splat-MOVER	RGB-D	2024	GPU
F3RM	RGB-D	2023	GPU

This integral is approximated using rasterisation techniques, where each Gaussian is splatted onto the 2D plane as a 2D Gaussian footprint. The rendering process is fully differentiable, enabling gradient-based optimisation of the Gaussian parameters during scene reconstruction.

In the reconstruction pipeline, given target views and their corresponding ground-truth images, the parameters $\{\mu_i, \Sigma_i, A_i\}_{i=1}^{N}$ of the Gaussians are optimised by minimising a rendering loss $\mathcal{L}$. For example, the Mean Squared Error (MSE) between the rendered image $\hat{I}$ and the ground-truth image I can be written as:

$$\mathcal{L} = \frac{1}{|\mathcal{I}|}\sum_{\mathbf{u}\in\mathcal{I}} (\mathcal{P}(\mathbf{u}) - I(\mathbf{u}))^2,$$

where $\mathcal{I}$ represents the set of pixel positions in the image. The gradients of this loss function with respect to the Gaussian parameters are computed, and the parameters are updated iteratively using gradient descent.

The covariance matrix Σ_i encodes the shape and orientation of each Gaussian, enabling it to adapt to local geometric features such as edges or surface curvatures. Additionally, the amplitude A_i controls the intensity or opacity of each Gaussian. The projection and blending of overlapping Gaussians during rasterisation ensure a smooth and visually coherent rendering.

Gaussian splatting's mathematical formulation as a mixture model over 3D space provides a probabilistic interpretation. It eliminates the need for explicit surface connectivity or iterative ray marching, as required in voxel grids or neural radiance fields (NeRFs), respectively. Instead, the closed-form Gaussian functions and their differentiability make Gaussian splatting computationally efficient and well-suited for large-scale scene reconstruction and rendering tasks. Papers covered in this section are provided in Table 4.

4.3. Applications

In robotic navigation, Chen et al. [93] used Splat-Plan and Splat-Loc for safe trajectory planning and robust pose estimation. Compared to traditional navigation planning methods [94,95] uses polytope corridor through the GSplat map to ensure safety by adhering to collision constraints. Within this corridor, it generates a Bézier curve trajectory for smooth navigation. Earlier works required semantic prior knowledge [96,97] for pose initialisation in navigation. In contrast, by leveraging the point-cloud nature of GSplat scenes, this approach enables global pose initialisation without prior knowledge and supports real-time pose localisation using only RGB images.

Similar to navigation object-centric semantic integration in [98] allows Gaussians with identical semantic labels to be initialised and updated collectively. In contrast to other methods that rely solely on image embeddings and lack geometric semantic representation [99], Gaussian splats integrate RGB details with geometric representation.

Table 5
Gaussian Splat in robotics : Results across different Datasets, Grasping Success rate is in percentage, and Task consists of slide block,put in drawer,drag stick,push buttons,stack blocks.

Dataset	Method	Reconstruction(PSNR)	Segmentation(mIoU)	Grasping(g)/Task(t)	ATE (↓)
KITTI	Street Gaussians [108]	25.79 (75% dataset)	58.81	–	–
	3DGS [92]	36.67	–	–	–
	HUGS [109]	25.42 (KITTI360/scene02)	85.64	–	–
TUM RGB-D	HF-SLAM [65]	22.60 (fr1/desk)	–	–	3.38 (fr1/desk)
	E-SLAM [110]	11.29	–	–	2.47
	Point-SLAM [87]	13.87	–	–	4.34
	DDN-SLAM [111]	–	–	–	0.9
	NeSLAM [112]	–	–	–	1.6
Replica	HF-SLAM [65]	35.74	–	–	0.34
	Gaussian-SLAM [113]	28.539	–	–	0.31
	Point SLAM [87]	21.3	–	–	0.52
	SplaTAM [60]	19.33	–	–	0.36
other	ManiGaussian(RL Bench) [114]	–	–	44.8(t)	–
	GNFactor(RL Bench) [115]	–	–	31.7(t)	–
	LSeg(GaussianGrasper data) [116]	–	26.4	26.7(g)	–
	Splat-MOVER(GraspAffordance-saucepot) [117]	–	–	100(g)	–
	F3RM (GraspAffordance-saucepot) [118]	–	–	30.0(g)	-

This combination enhances semantic quality and allows for rapid adjustments in response to the movements of robots and objects, facilitating real-time updates of the scene.

Recently, a novel representation that combines geometry, physics, and visual observations to improve robotic perception, planning, and control has been introduced through the incorporation of physical properties with Gaussian Splatting [100]. One key advantage of using Gaussian splatting is its ability to provide a continuous and differentiable representation of the visual state, which facilitates efficient rendering from any viewpoint. This is in contrast to traditional methods [101, 102] that may depend on discrete or less flexible representations, potentially resulting in less accurate or slower updates.

Instead of single tasks like mentioned above, multi-tasking robots [103] are using dynamic Gaussian splatting instead of dynamic NeRF [104]implicit methods, due to faster training, to model the propagation of diverse semantic features in the Gaussian embedding space, allowing for better comprehension of object interactions and scene-level spatiotemporal dynamics and for real-time performance.

Even in the application of SLAM [105], Gaussian splatting plays a key role, where Gaussian splat SLAM [106] can reconstruct tiny and even transparent objects, which are often challenging for traditional SLAM techniques. Also, the method achieves state-of-the-art results in novel view synthesis, surpassing traditional SLAM approaches.

In disaster relief scenarios, it is imperative for humans to maintain a safe distance while robots undertake navigation through hazardous environments to perform designated tasks. Given the high-stakes nature of these situations, rapid decision-making by human operators is essential, facilitated by a virtual representation of the environment that includes the robot's location. To optimise decision-making efficiency, buffering of scenes should be eliminated. Furthermore, as Gaussian Splatting has demonstrated superior performance in rendering speed compared to Neural Radiance Fields (NeRF) [12], it proves to be a more effective method for enhancing human–robot interactions in such contexts [107].

4.4. Evolution of Gaussian splatting in robotics

The evolution of Gaussian splatting techniques in 3D reconstruction and rendering has been marked by significant advancements, enhancing efficiency, quality, and applicability across various domains. The concept was first introduced by Westover in 1991, laying the groundwork for Gaussian-based volume rendering [119]. In 2023,[92] demonstrated a leap in computational efficiency, enabling high-quality real-time radiance field rendering.

Building upon previous work, [120] tackled challenges associated with view consistency by implementing a sorted splatting mechanism, which markedly improved stability during rendering processes. Simultaneously, [121] introduced a split algorithm focused on enhancing surface uniformity and adherence, thereby facilitating explicit editing and allowing for more accurate extraction of point clouds. The improved surface uniformity achieved through the split algorithm enables the production of high-fidelity maps, which will advance obstacle detection and path planning in the robotics field in the future.

Subsequent developments have sought to enhance the core methodology. For instance, the integration of surface normals into the rendering pipeline has been shown to boost surface detail estimation and visual quality [122]. Expanding on traditional Gaussian splatting, the replacement of Gaussian kernels with linear ones achieved faster computations and superior accuracy [123]. Addressing the issue of over-reconstruction, frequency-based techniques were incorporated to ensure higher fidelity in reconstructed scenes [124].

Additionally, a multi-scale approach was introduced to mitigate aliasing artifacts, ensuring smoother rendering [125]. Finally, the integration of spatial and angular Gaussians expanded the applicability of Gaussian splatting, achieving high-quality real-time relighting and view synthesis [126]. This is essential for tasks like object recognition and manipulation in dynamic environments for robots. Collectively, these works chart a path of continuous refinement, showcasing the adaptability and growing impact of Gaussian splatting in robotic applications.

4.5. Quantitative results

The results of Gaussian Splat techniques across various datasets—for tasks including reconstruction, segmentation, grasping success rates, and absolute trajectory error (ATE)—are presented in Table 5. For the KITTI dataset, methods like Street Gaussians achieve a reconstruction PSNR of 25.79 for 75% of the dataset, while HUGS records a higher segmentation performance (mIoU of 85.64) on specific scenes like KITTI360/scene02. Reconstruction quality (PSNR) improves with methods such as 3DGS (36.67), indicating advancements in Gaussian-based methods for street-level data (see Table 5).

In the TUM RGB-D dataset, HF-SLAM achieves a reconstruction PSNR of 22.60 for specific scenes (fr1/desk) but suffers from a higher ATE of 3.38. More advanced methods like NeSLAM achieve lower ATE values (1.6), indicating better trajectory estimation performance. Point-SLAM demonstrates a balance between reconstruction quality (PSNR of 13.87) and moderate ATE (4.34).

For the Replica dataset, HF-SLAM outperforms other methods in reconstruction with a PSNR of 35.74, followed by Gaussian-SLAM (28.539). The ATE values for Replica remain relatively low across

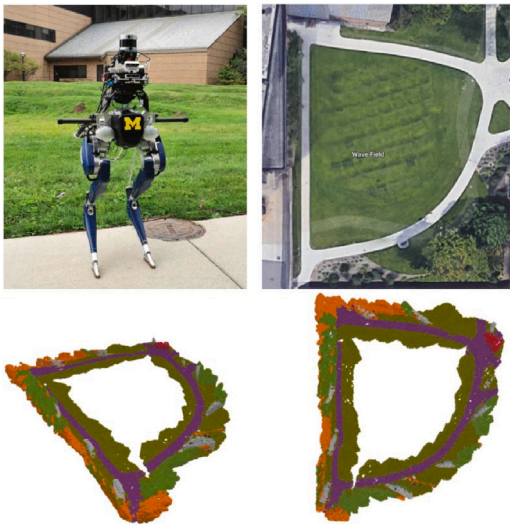


Fig. 2. Semantic 3D Reconstruction of a real-life sample park area provided by [127] using a Cassie Robot fitted with VLP-16 LiDAR and Intel RealSense D435 Camera. Top left: Cassie Blue with a custom torso mounted with an Intel RealSense depth camera, providing RGB images and point clouds in outdoor environments. Top right: Google satellite map of the Wave Field at the University of Michigan - North Campus. Bottom: 3D and 2D views of the S-BKI map. During navigation along the sidewalk, S-CSM produces discontinuous semantic maps from sparse sensor measurements, causing potential issues for the robot's planner. In contrast, the S-BKI model generates a continuous and smooth map, inferring labels for gaps from local correlations.

methods, such as HF-SLAM (0.34) and SplaTAM (0.36), showcasing effective trajectory estimation in simulated environments.

In the other datasets, specific tasks focus on grasping success rates using Gaussian-based methods. ManiGaussian achieves a grasping success rate of 44.8%, while Splat-MOVER excels with a 100% success rate. GNFactor also report high grasping rate 31.7% for specialised grasp affordance tasks.

Overall, Gaussian-based methods demonstrate strong performance across diverse datasets, with high reconstruction quality (PSNR) in KITTI and Replica, effective grasping success rates in specialised tasks, and promising trajectory estimation accuracy (low ATE) in TUM RGB-D and Replica datasets. These results highlight the versatility of Gaussian techniques in both visual and manipulation tasks.

5. Semantic 3D reconstruction

5.1. Overview

Semantic 3D Reconstruction/Labelling is a technique which is used in robotics to reconstruct and label 3D reconstructions to predict and detail what the object is perceived by giving the reconstruction a certain colour to distinguish the objects from the other types of objects. As shown in Fig. 2, provided by [127], it shows the separation between different parts of the scenery such as paths and vegetation. Semantic 3D reconstruction is increasingly important in robotics as it facilitates a deeper integration of perception and action. The technique aids in distinguishing and categorising objects within a scene by colour-coding them based on their semantic properties, as demonstrated in various studies [127–130]. This not only enhances the visual data but also adds a layer of informative context that is critical for autonomous operations in both indoor and outdoor settings. when it comes to modern research semantic 3D reconstruction is extended with the use of Gaussian-Splatting which works by representing each point in a point cloud with a Gaussian kernel, which spreads the point's influence over a local area to create a smooth surface. These kernels overlap and combine to form a continuous, detailed representation of the surface, allowing for efficient rendering and visualisation of complex datasets. SGS-SLAM [131] used Gaussian-Splatting for the first time to incorporate semantic features with geometry using multi-channel optimisation. A Spatially Consistent Feature Fusion model is used to avoid erroneous estimation of semantic maps using Gaussian-Splatting [132].

The analysed papers [127–130,133–138] provide insights into the use of semantic 3D reconstruction for both indoor and outdoor applications, highlighting how this technique enhances the understanding and interaction with the perceived scene by categorising and labelling different objects within the environment.

Indoor environments are reconstructed based on different types of objects in the room, such as chairs and tables, to label these objects for the robot to recognise and interact with. For instance, [128] focuses on labelling objects within a scene to position them for disinfection, though the actual disinfection process is not covered due to the lack of a robot in the study. In contrast, [134] uses 3D labelling and reconstruction to categorise objects within a space, incorporating both robotic and human perspectives to enhance accuracy. This approach, however, faces challenges in processing and labelling thin objects like chair legs.

In collaborative settings, [135] demonstrates a system where multiple tracked robots scan an area and semantically fuse all the scans to produce a global map. This method sparsely identifies objects and structures within its field of view, both indoors and outdoors, but does not compare the performance of single-robot mapping versus multi-robot mapping. Similarly, [129] creates semantic point cloud reconstructions from aerial drones, highlighting specific objects within the scene in a manner akin to [128].

In the realm of real-time dense reconstruction, [138] showcases a robotic rover equipped with an RGB-D camera to capture and reconstruct indoor scenes, providing a detailed view similar to the method employed by [134].

Outdoor environments within semantic 3D reconstruction allow robots to operate in real-life scenarios, reconstructing and analysing perceived objects. Using a mix of ground and aerial robots, semantic reconstruction is achieved both at ground level and in the air. Studies such as [127,130,133,135–137] employ multiple robots to reconstruct a global map of the surrounding area through simulations and real-life tests. These 3D semantic maps come in both sparse and dense forms. For example, [133,136] focus on sparse maps, while [127,130,137] produce dense maps using voxel grids for accurate environment reconstruction.

Semantic labelling is achieved in various forms, such as picking out key objects to navigate the scene around the robot [136], and using dense semantic maps to accurately reconstruct the perceived environment [127,137]. Some studies, like [130], use multiple robot configurations to sparsely semantically map out the environment in both simulation and real life, similar to the approaches in [133,135,136]. However, these methods produce incomplete sparse maps, which are not suitable for high-quality dense reconstructions.

A single robot can be employed to focus on scaffolding constructions within a building site, achieving sparse reconstruction but encountering difficulties in busy or chaotic environments [133]. In contrast, accurate dense semantic 3D reconstructions can be achieved using a vocalisation algorithm for scene reconstruction in both simulations and real-life tests. However, the short-range sensors attached to the robot result in missing middle sections of the reconstruction due to the limited range of the forward-facing camera [127]. LiDAR scans are used to produce semantic maps of both indoor and outdoor scenes, labelling objects within the trajectory to create a global 3D semantic map, with multiple robots often employed to reduce the time required for sparse reconstruction [130,135]. This approach is extended by using multiple robots, including a car and a drone, to identify objects within the scene and perform 3D reconstruction based on their positions relative to the device using SLAM. These maps are then augmented to create a global location map [136]. The potential of multiple robots

Table 6
Summary of main Semantic 3D reconstruction papers discussed in this section.

Paper	Sensors used	Year	Platform (CPU/GPU/FPGA)
RA-SLAM [128]	RGB-D Camera	2022	CPU + GPU
CrossSemantic3D [134]	RGB-D + Simulated Sensor	2023	CPU + GPU
S-BKI [127]	LiDAR + Semantic Annotations	2020	CPU
RTSDM [129]	RGB-D Camera	2022	CPU + GPU
HD-CCSOM [137]	LiDAR + Semantic Labels	2022	CPU
RD3DSM [138]	RGB-D Camera	2023	CPU + GPU
Kimera-Multi [130]	Stereo Cameras (Simulated)	2021	CPU + GPU
3DSemanticRobotDog [133]	RGB-D + Inertial Sensors	2022	CPU + GPU
MR-SLAM [136]	Stereo + UAV Sensors	2021	CPU

equipped with laser scanners is also demonstrated to enhance the accuracy and efficiency of semantic 3D reconstructions in selected outdoor scenes [137].

These studies highlight different methods and challenges in achieving semantic 3D reconstruction in outdoor environments, demonstrating the capabilities and limitations of modern SLAM technologies in various real-life scenarios.

The collection of studies on Semantic 3D reconstruction offers nuanced insights into both indoor and outdoor applications, each underlining specific advantages and facing distinct challenges. For indoor scenarios, works like those by [128,134] excel in environments requiring precise object recognition, though they struggle with complex geometries and detailed features like thin chair legs. Conversely, outdoor applications described by [127,130] manage dynamic conditions well, though they demand significant computational resources. A recurring theme across the papers is the use of both sparse and dense reconstruction techniques; sparse methods [133] are faster and suited for rapid navigation, while dense methods [127] provide more detail at a greater computational cost. Additionally, the employment of multiple robots enhances data collection but also introduces complexities in data integration, often without a proportional increase in performance over single robot systems, highlighting an area ripe for further development. All the papers in the sections are provided in Table 6.

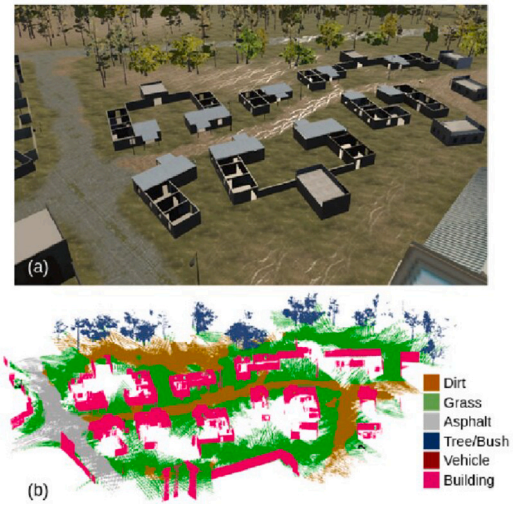


Fig. 3. Semantic 3D Labelling as shown in [130]. (a) Camp scene created using the Unity-based DCIST multirobot simulator. (b) Dense metric-semantic 3D mesh model produced by Kimera-Multi with the collaboration of three robots.

5.2. Processes of 3D semantic reconstruction

5.2.1. Overview

Semantic 3D reconstruction has two constituent parts, Semantic labelling and 3D reconstruction. The processes involved in 3D Semantic Reconstruction are crucial for advancing the capabilities of robotic systems, enabling them to interact with their environments more effectively. This technology integrates detailed 3D mapping with semantic understanding, which allows robots to not only navigate through but also understand and interact with their surroundings intelligently. It enhances robotic perception by categorising objects based on their significance and context, which is essential for applications ranging from autonomous navigation to complex task execution. By providing a richer, more informative representation of the environment, 3D Semantic Reconstruction helps in developing more autonomous, efficient, and adaptable robotic systems. Overall, this process is fundamental for the evolution of robotics, pushing the boundaries towards creating machines that can operate independently in a variety of settings. Semantic labelling labels the objects in the frame of view and produces a colour to indicate the type of object, as stated in [127,129,130,134–138]. Also this method can be used to clarify a specific type of object, as mentioned in [128,129]. Semantic labelling is trained on data sets such as the KITTI [139] and Semantic3D [140] data sets to label objects within the frame of view by colour to produce a 2D colour labelled map of a single frame image of the scene. After a video has been created of the semantic 2D labelled pictures, this is then applied to reconstructions of scenes with the applied object colour predicted, to show the dimensions of certain objects within the reconstruction, as shown in Fig. 2. These reconstructions can be based on locations of objects in space [136], highlighting specific objects within a scene [128,129] or producing sparse [130,133,135] and dense [127,134,137,138] 3D semantically labelled reconstructions. As shown in Fig. 2, objects are detailed in the reconstruction with the appropriate colour to allow the robot to understand what type of object in the scene it is.

5.2.2. 3D semantic labelling

As briefly stated above, 3D Semantic labelling is the use of categorising objects from a data set using colour to visibly indicate a particular object by using 2D labelling techniques. Then using SLAM and multiple 2D semantic labelled maps, the whole scene is represented by the objects' colour and a reconstruction of the scene. Methods show that reconstructions of scene come in two forms, sparse [130,133,135,136] and dense [127–129,134,137,138] 3D reconstructions.

As shown in Fig. 3, sparse reconstruction of the scene typically consists of a sparse point cloud, where the frame-by-frame 2D semantic labelled scenes are augmented with the reconstruction to assign each point in the point cloud a colour corresponding to the object it represents. This method is exemplified in various studies that produce semantic reconstructions with sparsely coloured point clouds of simulated scenes using their proposed SLAM methods, showcasing incomplete but accurately labelled objects.

One study produces a semantic reconstruction similar to another, demonstrating sparsely coloured reconstructions of simulated scenes with accurate labelling of objects within the reconstruction [130,133]. This approach provides sufficient information for robotic systems to understand and navigate their surroundings. Recently, SlideSLAM [141] used a hierarchical metric-semantic representation of the environment, including high-level sparse semantic maps and low-level voxel maps to explore indoor as well as outdoor scenes including forests. Another study further demonstrates this method in a chaotic environment, such

Fig. 4. 3D TSDF Semantic Reconstruction shown in [128]. Comparison of 3D semantic reconstructions for ScanNet scene 0665_00, highlighting high-touch surfaces. High-touch objects like tables and chairs are marked in red, while other surfaces are in green. Our system's reconstruction (right) successfully identifies most high-touch areas present in the ground-truth reconstruction (left). (Best viewed in colour).

as a building site, by reconstructing scenes and locating scaffolding frames with specific colours, akin to the method proposed in another paper [128,133].

In another approach, LiDAR is utilised to produce similar-style scans of the scenes, demonstrating how multiple robots can enhance the reliability and coverage of scans [135]. This method, similar to that used in other studies, shows the benefits of employing multiple robots for broader area coverage and more reliable scans [130,135].

A different approach focuses on detailing only the objects within the scene, creating a map that represents objects as small squares located at their positions relative to the capturing device using SLAM. This results in a globalised map showing the relationship of objects from the merged scenes captured, enhancing the overall understanding of the environment [136].

These studies illustrate various techniques and applications of sparse semantic reconstruction, demonstrating the potential for enhancing robotic navigation and interaction with real-world environments [128, 130,133,135,136]. For dense reconstruction of these scenes, methods such as marching cube algorithms and dense point maps are utilised to produce detailed meshes of the environment. For instance, vocalisation methods are used to reconstruct the scene around the robot, with multiple captured frames colour-coded according to the reconstructed objects [127]. Another approach uses a truncated signed distance function (TSDF) method to highlight surfaces in the scene with a specific colour, enabling a robot to navigate and clean selected surfaces [128]. A similar method combines robot and human-based views to fuse their perspectives and produce a more accurate reconstruction, which is then semantically labelled [134]. In a different method, particular objects within a scene are highlighted with specific colours on the global reconstruction from the captured images [129]. Simulated reconstruction methods using multiple robots create dense point cloud reconstructions, labelling objects at specific points in the cloud [137]. Gaussian-Splatting which can be considered as cluster of local points to create a single splat is also used in dense reconstruction semantic labelling [142]. Real-time methods combine human and robot views to show reconstruction and labelling capabilities from different angles using public datasets for categorisation [137]. These dense reconstruction techniques enhance the detail and accuracy of 3D scene reconstructions, demonstrating their potential for improving robotic navigation and interaction with environments.

5.2.3. 3D semantic reconstruction techniques

Reconstruction techniques used on robotic systems to categorise objects within the perceived scene in a 3D form come in particular forms, using methods such as TSDF [128,134], as shown in Fig. 4, and voxelisation algorithms [127], as shown in Fig. 2, to produce dense reconstructions. Other methods use sparse techniques such as the use of SLAM algorithms that use the raw depth data to produce a low-cost point cloud map of the scenes captured on the robots used.

Semantic 3D reconstruction involves two processes: 2D Semantic labelling, which categorises objects perceived in a video, and a 3D reconstruction method that fuses these labels onto a mapped reconstruction. The TSDF reconstruction method is used to augment Visual SLAM and Semantic Segmentation, providing depth visualisation and segmenting surfaces to produce a 3D Semantic map of surfaces [128]. This method does not test the algorithm on a real robotic system, leaving its practical application uncertain. Another approach merges semantic maps based on object types and uses two different viewpoints to increase accuracy [134]. A voxelised map of the scene is produced using a Cassie robot fitted with a LiDAR sensor and stereo cameras, capturing 2D semantic maps of the environment and merging them with a 3D voxel map, projecting colours from the semantic map onto the reconstruction [127]. Scenes are reconstructed based on multiple images, creating a dense point cloud with selective semantic object detection implemented into the scene [129]. Simulated laser scans and reconstructions of the environment from multiple robots employ the Octomap method to produce dense reconstructions [137]. This method is also used to reconstruct semantic scenes in real-time on a robotic rover [138]. These approaches showcase various techniques in Semantic 3D reconstruction, highlighting the potential for detailed and accurate environmental mapping in both simulated and real-world scenarios.

Sparse reconstructions shown in [130,133,135,136] show the use of point clouds and laser scans to produce semantic 3D reconstructions of the environment perceived. Point clouds are shown in [130,133], showing a sparse reconstruction of the objects within the scene with each point being categorised as a particular object. Specifically, with the use of a stereo camera and a LiDAR sensor, [133] shows scans of a building site with categorisation of the objects around it, with a complete reconstruction of scaffolding frames in the scenes provided. Object categorisation of all objects within the scene is demonstrated using a reconstruction method that categorises objects in a similar form to previous studies. This method provides detailed labelling and categorisation of objects, enhancing the robot's ability to understand and interact with its environment [130,133]. These approaches underline the effectiveness of semantic labelling in creating comprehensive 3D maps for robotic applications. Using a reconstruction that only uses a LiDAR sensor and has a similar semantic reconstruction method as [130,135] uses multiple robots to produce fused laser scans of outdoor and indoor environments with object categorisation shown in the laser scans provided. Using semantic techniques similar to [130, 135,136] produces a 3D map of locations of objects related to the device used and produces a webbed map with circles illustrating the objects perceived.

The exploration of Semantic 3D Reconstruction techniques across various studies illustrates significant advancements in enhancing robotic capabilities for environment interaction and object recognition. Techniques like TSDF and voxelisation [127,128] offer detailed environmental mappings crucial for precise robotic navigation, while the integration of multi-modal data from sources like LiDAR and visual inputs [135,137] enriches the semantic mapping process, aiding robust environmental representations. Despite these benefits, challenges persist, particularly with the high computational demands and the complexity of data integration seen in studies like [130,136], which complicate real-time applications. Moreover, maintaining accuracy in dynamic environments [133,134] remains problematic due to the transient nature of objects and variable conditions. These findings underscore the need for ongoing advancements in computational efficiency, data processing speed, and methods for integrating diverse data sources to leverage the full potential of Semantic 3D Reconstruction in practical robotic applications.

Semantic 3D reconstruction techniques in robotic systems, several studies provide compelling use cases. For instance, the work by Agnew et al. [2] introduces an Amodal 3D reconstruction system

Table 7
Quantitative results of various semantic 3D reconstruction approaches, categorised into dense and sparse methods. Metrics include Intersection over Union (IoU), Mean Accuracy, Collection Runtime, and Mean Absolute Trajectory Error (ATE). The table compares different implementations based on these metrics to highlight their performance and effectiveness.

Papers	Dataset	IoU (%)	Mean Acc. (%)	Coll. Runtime (ms)	Mean ATE (m)
Dense Semantic Reconstruction Approaches					
RA-SLAM [128]	ScanNet [144]	51.2	–	–	–
CrossSemantic3D [134]	Habitat Simulation Framework [134]	–	73.8 (Avg of Table 4)	–	–
S-BKI [127]	KITTI-semantic [145]	57.1 (table III)	–	68.17	–
RTSDM [129]	self-built [129]	85.1(Avg of Table 3)	–	111.5 (max)	0.11345
HD-CCSOM [137]	KITTI-Semantic(seq04) [145]	62.58	–	–	–
RD3DSM [138]	SUN-RGB-D [146]	32.32(Avg of Table 2)	–	47.83 (max)	–
Sparse Semantic Reconstruction Approaches					
Kimera-Multi [130]	Simulated Data (City-Merged) [130]	–	83.93 (Table III)	3125 (RBCD (ES))	1.54025 (RBCD (ES))
3DSemanticRobotDog [133]	self-built (Yonsei University) [133]	91.3 (Table 2 -max)	–	–	–
MR-SLAM [136]	KITTI [147]	–	–	248 (Car and UAV)	2.12 (Car and UAV)

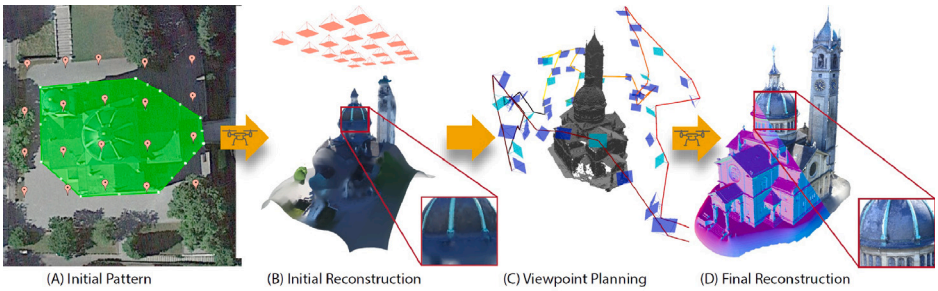


Fig. 5. An end-to-end system has been developed for 3D reconstruction of large-scale scenes using commercially available quadrotors [148]. (A) The process begins with a user defining a region of interest (green) on a map interface and specifying a flight pattern of viewpoints (orange) at a safe altitude. (B) The quadrotor follows this predefined pattern, capturing images that are processed to create an initial reconstruction and occupancy map. (C) Next, a viewpoint path is computed to maximise the observation of unknown spaces, adhering to the constraints of a specifically designed camera model. This path is restricted to known free space, allowing for fully autonomous execution of the trajectory. (D) The final step involves processing the newly captured images to produce a high-quality reconstruction of the region of interest, effectively capturing concave areas and fine geometric details.

that significantly enhances robotic manipulation tasks in cluttered environments. Their experiments demonstrate a 42% improvement in manipulation success rates for previously unseen objects, showcasing the effectiveness of integrating object stability and connectivity priors in reconstruction. Additionally, a study utilising a Cassie robot equipped with LiDAR and stereo cameras successfully captures 2D semantic maps and merges them with 3D voxel maps, resulting in detailed environmental reconstructions that facilitate better interaction with the surroundings [143]. These examples underscore the practical applications and advancements in semantic 3D reconstruction methods, highlighting their impact on enhancing robotic capabilities for environment interaction and object recognition.

5.2.4. Quantitative results

The quantitative results of [127–130,133–138] are discussed in the respective experimental results sections. However, due to inconsistencies between the reported data and the actual experimental results, [135] has been excluded from the comparison. A summary of these results is provided in Table 7.The lack of evidence was present in references due to not being present within the above-mentioned papers themselves. With the available results in the Intersection over Union (IoU) section, some papers [128,129,133,137,138] showed to measure this value in two different aspects, by the accuracy of the semantic label or the reconstruction accuracy of the scene [127]. As these two are based on the label and reconstruction quality of an object related to the image taken, these two values can be placed in the same column as it is about the quality of the reconstruction. Similar to IoU values, the overall accuracy [130,134] and the confidence level [135] of reconstructions created by the methods stated are measured within Table 7 to provide information on the accuracy of reconstructions related to others. Runtime for methods with the data applicable [127, 129,130,136,138] shows how long the algorithm takes to compute the reconstruction that is shown in the papers. [127] shows a relatively quick time due to tests being conducted in synthetic scenes, similar to [129]. To deliver reconstructions in real-time, [138] sacrifices accuracy over performance, to deliver such small reconstruction processing time. With [130], a much larger scene is reconstructed and so takes a longer time to compute. With this, the error is stated on which it shows again the error of the reconstruction that is created based on the scene shown. Though [130] has a larger running time, the paper shows a lower error than [136] on their most computational method.

6. Multi-view 3D reconstruction

6.1. Overview

With the methods stated above, the use of multiple viewpoints on the object that the algorithm is trying to reconstruct normally improves the general quality of the reconstruction. This section provides methods that use multiple viewpoints from multiple devices or angles to produce high-quality reconstructions. As shown in Fig. 5, from [148], it shows a reconstruction of a building with the use of an aerial drone fitted with a stereo camera. Most of the papers [34,148–150,152] shown use a trajectory to produce multiple views of the object or building perceived by using Structure from Motion (SfM) algorithms. Other methods reconstruct objects from multiple images taken from different positions of the object whilst the capturing device is in a fixed position [157] or moved on a robotic arm [151,153,158]. These methods, as represented by the referenced papers, used different reconstruction techniques, such as multi-view stereo (MVS), to reconstruct the object or surface perceived. All the papers in this section are provided in Table 8.

6.2. Processes of multi-view 3D reconstruction

6.2.1. Overview

Multiple view reconstructions are created from multiple frames that are taken from different sides of the object and are reconstructed and

Table 8
Summary of main Multi-View 3D reconstruction paper discussed in this section.

Paper	Sensors used	Year	Platform (CPU/GPU/FPGA)
Sequential-Reconstruction [149]	RGB-D + Depth Sensor	2022	CPU + GPU
Plan3D [148]	RGB-D (Synthetic)	2018	CPU
NodeSLAM [150]	RGB-D Camera	2020	CPU + GPU
ROBI [151]	RGB Camera + Depth Sensor	2021	CPU + GPU
MVS-Path-Planning [152]	Simulated Stereo Cameras	2021	CPU + GPU
Reconstruction in Welding [153]	Industrial Camera + Depth Sensor	2023	CPU + GPU
DUSt3R [154]	Multi-view RGB Cameras (DTU)	2024	GPU
Detector-Free SfM [155]	Multi-view RGB (ETH3D)	2024	CPU + GPU
VGGSfM [156]	RGB Camera (ETH3D Benchmark)	2024	GPU

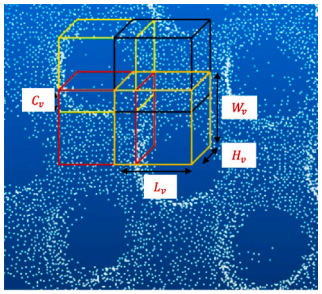


Fig. 6. This figure illustrates the definition of voxels within the tubesheet point cloud [153]. The image depicts multiple overlapping voxel grids, with dimensions L_v, W_v, and H_v representing the length, width, and height of each voxel, respectively. The central voxel C_v is highlighted to show its position relative to the other voxels. This visualisation aids in understanding the spatial partitioning and organisation of the point cloud data within the tubesheet.

fused on a global map using GPS data [34]. These scenes and objects are captured through the use of stereo [34,151,158] and monocular cameras [148–150,152,153,157] on which use specific reconstruction methods such as the Iterative Closest Point (ICP) algorithm [151]. Then, a global reconstruction map is created, such as an occupancy map [148], to show all captured angle reconstructions on one map. This section splits this method into two sections: the types of reconstruction techniques used and how they fuse the reconstructions together in the papers proposed.

6.2.2. Types of reconstruction techniques used in multi-view reconstruction

Multiple view reconstructions on the proposed robotic systems are based on the use of point clouds to provide a smaller cost factor for performance on computerised devices provided than other reconstruction methods. Point clouds are created with the use of stereo [34,148,151,152,158] and monocular cameras [149,150,153,157] to provide a cost-effective capturing solution, on which after are correlated and the use of triangulation [149,152,153,158] is used to fuse the camera frames together or using more traditional methods such as developing a depth map to develop a mesh [34,148,151,152] and use GPS data to fuse the reconstructions together. As shown in Fig. 6, Wang et al. [153] use methods similar to [149] to produce a point cloud of an example tube-sheet using a monocular camera. As point clouds are a relatively cost-effective method to reconstruct objects, [149] shows the use of an aerial robotic system to capture a selective environment. Also, with the use of drones in a selective environment, [148] uses stereo cameras to develop a highly detailed reconstruction from multiple aerial views of buildings using occupancy maps. Using stereo cameras similar to the ones used in [148] and the use of a similar point cloud reconstruction method in [153,158] uses stereo cameras to produce a 3D point cloud of a model satellite to allow for a more precise method of manipulation of the robot arms compared to the perceived object. Using monocular views of objects and occupancy maps shown in [148,150] produces high-quality reconstructions of objects by fusing multiple views of the objects and using a custom CAD object dataset to help improve future robot manipulation tasks similar to [158]. Using stereo cameras, [151] produces a dataset of complex metallic objects within a tight space to help with robotic manipulation [150,158], and demonstrates this within the paper. Similar to [148] and using a reconstruction method similar to [152,153,158] produces large-scale and accurate reconstructions using an aerial robotic vehicle. Using one point of view to reconstruct the surface of an object [153,157] tracks the object from one point of view to produce a multi-view reconstruction of the object as the object is moved within the selective frames. Within an agricultural setting, [34] shows the use of methods shown in [148,150] to produce dense reconstructions of the field around the robotic vehicle using stereo cameras. Multi sensor fusion also used by harmoniously blending LiDAR-inertial odometry (LIO), visual-inertial odometry (VIO), and sophisticated Inertial Measurement Unit (IMU) specially to create better accurate model in variable elements [159]. A real time reconstruction specially in mobile robots was achieved using MVS-SLAM model which only used RGBD camera, to enhance multi view geometry [160].

6.2.3. Types of fused reconstruction techniques used in multi-view reconstruction

In the papers proposed, fusion techniques that are used are based on the camera used for the reconstruction of the scene or objects. The main two methods are Structure from Motion (SfM) [148,149], TSDF [151,157] Fusion, and average surfel models [152]. SfM takes multiple views of an object, which have no position related to each other, and feature matches them to produce a reconstruction. This method is used by monocular cameras as it does not need to rely on reliable concrete distances between the lenses. TSDF Fusion takes multiple depth maps from multiple views perceived from stereo cameras to produce high-quality reconstructions as shown in [148,151]. Similar to TSDF Fusion, [34] uses a similar approach to TSDF Fusion, but uses Kalman filtering to predict reconstructed views. Average surfel models are multiple surfel models which have been averaged out to produce a more accurate model. Others use other custom methods that involve feature matching to fuse multiple point clouds together [153], and a mixture of cameras and methods to produce a more accurate point cloud representation of an object [158]. SLAM is closely tied with this method as the use of multiple views is essential for reliable trajectory planning for the manipulation of objects, as shown in [150]. As most of the previous methods rely on scene specific reconstruction IBD-SLAM [161] a scene-generalised model by using a mage-Based Depth Fusion. Recently, DUSt3R [154] introduces a novel approach to 3D reconstruction by eliminating the need for known camera parameters or poses. Traditional Multi-View Stereo (MVS) methods often rely on precise camera calibration, which can be cumbersome and error-prone. DUSt3R circumvents this by directly regressing 3D point maps from image pairs, allowing for dense reconstruction without prior camera information. This is achieved through a transformer-based architecture that leverages powerful pre-trained models. As traditional Structure-from-Motion [155] (SfM) pipelines rely heavily on detecting and matching keypoints across images, which can be challenging in texture-poor scenes. The Detector-Free SfM framework addresses this limitation by eliminating the need for explicit keypoint detection.

Table 9
Quantitative results of various multi-view 3D reconstruction approaches. Metrics include F-Score, Average Accuracy, and Error Rate. The table provides a comparison of different reconstruction methods based on these metrics, highlighting their performance and accuracy.

Papers	Dataset	F-Score (%)	Avg. Accuracy (m)	Error rate
Sequential-Reconstruction [149]	self built (ref sec 4.1) [149]	–	0.1867 (Avg of Table 4)	–
Plan3D [148]	Synthetic Scenes (ref sec 5.1) [148]	79.22 (Table 1 - Ground Truth Depth Map)	–	–
NodeSLAM [150]	DVR [162]	–	0.03484 (3-views)	–
ROBI [151]	ROBI dataset [151]	–	–	0.5775 (Distance Error (Zigzag))
MVS-Path-Planning [152]	RotorS simulation [163]	84.48 (Avg of Table II)	–	–
Reconstruction in Welding [153]	self-built [153]	–	–	9.03 (%)

Table 10
Quantitative evaluation of DUSt3R, Detector-Free SfM, and VGGSfM across various 3D reconstruction metrics. A dash (–) indicates that the specific metric was not reported in the respective paper.

Method	Dataset	RRE (°)	RTE (mm)	AUC@10° (%)	Accuracy (mm)	Completeness (mm)	Avg. Distance (mm)
DUSt3R [154]	DTU [164]	–	–	–	2.7	0.8	1.7
Detector-Free SfM [155]	ETH3D [165]	0.91	4.3	–	–	–	–
VGGSfM [156]	ETH3D [165]	–	–	75.39	–	–	–

Instead, it utilises detector-free matchers to establish correspondences directly. The framework reconstructs a coarse SfM model from quantised matches and then refines it through an iterative process involving attention-based multi-view matching and geometry refinement modules. This approach enhances robustness in challenging scenarios and has demonstrated superior performance over traditional methods, particularly in scenes with low texture. Additionally, VGGSfM [156] represents a shift towards end-to-end differentiable SfM pipelines. Unlike traditional methods that separate keypoint detection, matching, and bundle adjustment, VGGSfM integrates these components into a fully differentiable framework. It begins by extracting reliable pixel-accurate tracks using deep 2D point tracking, eliminating the need for chaining pairwise matches. Camera parameters are recovered simultaneously based on image and track features, and a differentiable bundle adjustment layer refines the reconstruction. This holistic approach simplifies the pipeline, reduces error accumulation, and enables joint optimisation of all components.

6.3. Quantitative results

The results from the proposed papers on multi-view reconstruction are discussed, though [34,157,158] have been excluded due to insufficient evidence regarding object reconstruction and related validation. While Chamfer Distance is reported in [157], it is the only work among the surveyed papers to include this metric. A summary of the included results is presented in Table 9. For the result of F-Score, [152] shows to have a higher F-Score than [148] on which each paper is reconstructing buildings. The average reconstruction accuracy shown in [149,150], shows that [150] has a better accuracy due to reconstructing multiple smaller objects on a flat surface than reconstructing a scene in [149]. The error rate, shown in [151,153], shows two different error rates, on which [153] shows an overall error rate in percentage for the reconstructions in the paper, and [151] shows a distance error for the difference from the actual size of the object and the reconstruction created.

The comparative evaluation of Table 10 highlights the performance of three prominent 3D reconstruction methods: DUSt3R, Detector-Free SfM, and VGGSfM. DUSt3R demonstrates strong capabilities in 3D reconstruction quality, achieving an accuracy of 2.7 mm, completeness of 0.8 mm, and an average distance of 1.7 mm on the DTU dataset. Detector-Free SfM excels in camera pose estimation, recording a Relative Rotation Error (RRE) of 0.91° and a Relative Translation Error (RTE) of 4.3 mm on the ETH3D dataset. VGGSfM achieves a notable Area Under the Curve (AUC) of 75.39

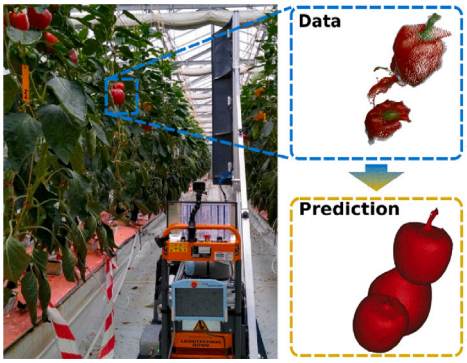


Fig. 7. Amodal reconstruction of fruit in a greenhouse environment shown in [166]. The left image depicts our robot monitoring a sweet pepper greenhouse near Bonn, Germany. Using an RGB-D frame (top right), this method can effectively complete and reconstruct the 3D shape of the fruits (bottom right).

7. Amodal 3D reconstruction

7.1. Overview

Amodal reconstruction is a band of 3D reconstruction which takes the perceived view and reconstructs it, like other previous methods, but then uses a Neural Network to complete the object related to a database of objects. In Fig. 7, Magistri et al. [166] use a database which they produce and use in this paper of fruit and vegetables to produce suitable quality reconstructions of fruit and vegetables, as shown above. Mainly this method is to improve the manipulation of objects by robots and so most papers in this area detail this [2,167–170,172–174]. The way that the object is reconstructed is different as some use point clouds to allow it to be less computationally heavy [169,170,172–174], which reduces the quality of the reconstruction, and some use TSDF mesh fusion which is computationally heavy but provides an accurate reconstruction of the object.

In this method, a dataset is used on a neural network to complete the reconstruction perceived by the robot, which requires object-based datasets such as YCB [175] or ShapeNet [176]. With this, it produces a likely representation of the object so a robot can manipulate the object correctly without dropping or damaging it. As shown in [167], the dataset YCB [175] is used to predict the full reconstruction of the object to allow it to be manipulated. This is similar to the other proposed papers that use their own modified datasets to allow for certain objects that it has been tailored to, to be reconstructed to the best of their ability. But, the previously stated methods when applied in real environment lacks physical realism, to tackle this ARM [177]

Table 11
Summary of main Amodal 3D reconstruction Papers used in this section.

Papers	Sensors used	Year	Platform (CPU/GPU/FPGA)
DSR-Net [167]	RGB-D Camera (self-built)	2020	GPU
ARM [2]	Synthetic Rendered RGB-D (ShapeNet)	2021	GPU
DDGC [168]	RGB-D Camera (Real Setup)	2021	CPU + GPU
kPAM-SC [169]	RGB-D Camera (self-built)	2021	CPU + GPU
Shape-Completion [170]	RGB-D Dataset (from [171])	2020	CPU
UOIS [172]	Real RGB-D (self-built)	2021	GPU
S4G [173]	Real RGB-D + synthetic	2020	GPU
3DSGrasp [174]	RGB-D	2023	GPU

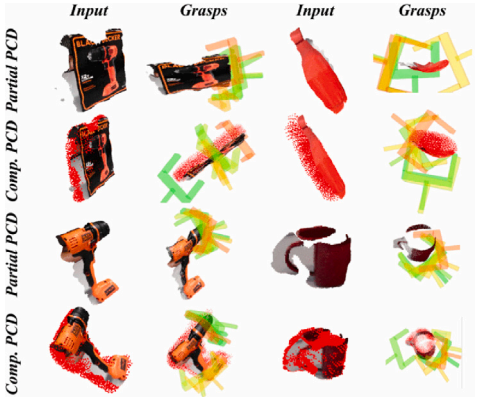


Fig. 8. Amodal reconstruction of objects for manipulation as shown in [174]. Qualitative results showcasing grasp proposals generated on both partial and completed point cloud data (PCD) for four objects using the PCD completion network. The partial PCDs are captured using a real sensor mounted on the Kinova robotic arm. Each candidate grasp pose, produced by the Grasp Pose Detection (GPD) algorithm, is colour-coded from green to red, representing scores from high to low.

uses priors over the physical properties to increase the physical realism. Specifically they incorporates priors on connectivity and objects. All the papers used in this section are provided in Table 11.

7.2. Processes of Amodal 3D reconstruction

Amodal reconstruction has two parts: the perceived reconstruction, which reconstructs the objects in a point cloud [166,169,170,172–174] or a dense model [2,167,168], and then shape completion, which takes the perceived reconstruction and completes the shape of the reconstruction with the use of a database of objects in a neural network [2,166–168,172–174] or other methods [169,170]. This section splits the processes stated above, to go into detail about what reconstruction methods have been used for the proposed papers, and how each proposed paper completes their reconstruction.

7.2.1. Reconstruction methods

Similar to other methods of 3D reconstruction, the use of sparse [166,170,172] and dense [169,173,174] point cloud methods are used as well as TDSF fused models [167,168] and voxel maps [2] are used to produce the perceived models. Two reconstructions of the object take place, one for the perceived object from the camera which gives an incomplete shape, and then the shape completion algorithm reconstructs the incomplete part of the object to give a complete reconstructed object. Most of these proposed papers use the same approach on reconstructing the perceived object, in point clouds [169,170,172,173] or TDSF fusion methods [167] or voxel maps [2], but some use a mixture of methods to show the difference between the shape completion reconstruction and the original perceived reconstruction [166,168,174]. For example, as shown in Fig. 8, Mohammed et al. [174] uses a mixture of coloured dense point cloud for the perceived reconstruction and an uncoloured sparse point cloud for the shape completed reconstruction. On this point, [168] shows the use of a dense point cloud for the perceived reconstruction similar to [174], but then uses a TDSF fused model to show the shape of a completed object. As well, as shown in Fig. 7, the use of a dense point cloud is used for reconstruction of the perceived view from the robot of the plant, and then similar to [168], produces a full object reconstruction of fruit using TDSF reconstruction. The benefit of using a point cloud is that it is less computationally than meshed methods like TDSF as there are fewer processes for the device to handle when analysing the object. With this, the data for the TDSF can be computed on a bigger device than the device perceiving the object as the data from the point cloud can be sent across a communication medium.

7.2.2. Shape completion

Shape completion involves using a perceived reconstruction, as stated above, to estimate the completion of a reconstructed object to improve object manipulation. The shape completion of an object typically employs a neural network, which is trained on a database of objects to estimate the missing parts of incomplete reconstructions [2,166–169,172–174,178], while other custom methods are also used to produce the missing reconstruction [170].

Regarding neural networks, the papers diverge in whether the network focuses more on predicting object grasping or on improving object reconstruction for better manipulation [168,169,173,174].

In grasp-oriented research, predicting grasp positions from a database of objects and viewed objects helps improve object manipulation [173]. One method uses a perceived dense point cloud reconstruction and an estimated sparse reconstruction of the missing object parts to predict grasp positions [174]. Another method estimates grasping on a meshed TDSF model of the object from an initial perceived point cloud reconstruction, which provides a more traditional approach [168]. Shape-completed objects are also used to predict key gripping points [169].

In reconstruction-oriented research, voxelised shape-completed reconstructions of objects use a TDSF perceived model, with the network improving based on previously provided objects [167]. Achieving a completed voxelised reconstruction of an object from a perceived view uses a network style similar to those used for grasp prediction [2]. Enhancing the density of the point cloud of an object improves manipulability outcomes from the reconstruction [172]. Both methods highlight accuracy issues in shape completion, suggesting that a reference database like YCB could be useful [175].

To address these issues, an algorithm utilises a custom reference database filled with reconstructed fruits and vegetables to improve the shape of completed objects [166]. This involves inputting a point cloud of the vegetable and shaping it into a meshed TDSF model. Other methods include a symmetry-aligned algorithm as depicted in Fig. 7 that predicts the symmetry line from the camera's position relative to the fruit, although it encounters challenges in accurately reconstructing the point cloud [170]. These approaches demonstrate various techniques for enhancing 3D reconstruction and grasp prediction in robotics.

Table 12
Quantitative results of various Amodal 3D reconstruction approaches. Metrics include Success Grasp Rate, Intersection over Union (IoU), and Error. The table compares voxel/mesh and point cloud shape completion reconstruction methods, highlighting their performance and accuracy.

Papers	Dataset	Success grasp (%)	IoU (%)	Error (cm)
Voxel/Mesh Shape Completed Reconstruction Approaches				
DSR-Net [167]	self-built (ref sec 4) [167]	–	62	0.048 (MSE)
ARM [2]	ShapeNet [176]	42	–	–
DDGC [168]	DDGC dataset [168]	40	–	–
Point Cloud Shape Completed Reconstruction Approaches				
kPAM-SC [169]	self-built (ref sec V-A) [169]	96	–	–
Shape-Completion [170]	Data obtained from [171]	–	61	–
UOIS [172]	self-built [172]	94.7	–	0.508 (MSE)
S4G [173]	S4G dataset [173]	92.5	–	–
3DSGrasp [174]	YCB [179]	76	–	–

7.3. Quantitative results

The quantitative results from the proposed robotics papers are presented, while other metrics—such as F-Score and Chamfer Distance (CD)—have been excluded due to their incompatibility with the evaluation criteria used in the remaining studies. These results are summarised in Table 12. CD of 0.92 has been shown in [174] and an F-Score of 69.43 from [166], which is the only data from the paper. The most common type of result that was quantitatively shown in the proposed papers was the percentage of successful grasps. The highest percentage of successful grasps was reported by [169], which shows a success rate of 96%, albeit tested on a limited number of objects. Conversely, papers like [2,168] reported lower percentages of successful grasps, indicating that voxel usage without refinement and cluttered scenes can lead to inaccurate reconstruction and shape completion for grasping.

To assess the quality of reconstruction, IoU results are shown in [167,170] to improve grasp performance, with similar findings between them. To monitor the trajectory of the arm picking up objects, Mean Squared Error (MSE) is used to enhance arm movement for object retrieval, as demonstrated in [167,172].

8. NeRF-based 3D reconstruction

8.1. Pose estimation and localisation

Recent advancements in Neural Radiance Fields (NeRF) have led to significant improvements in 6D pose estimation and localisation for robotics applications. NeRF-Pose [180] introduces a first-reconstruct-then-regress approach for weakly-supervised 6D object pose estimation, leveraging NeRF's ability to generate high-quality 3D representations from 2D images. This method demonstrates improved accuracy and robustness in scenarios with limited labelled data, making it valuable for robotic manipulation tasks.

NeRF-Loc [181] presents an innovative approach to robot localisation using NeRFs, achieving highly accurate 3D localisation by leveraging the implicit scene representation provided by NeRFs. This system shows superior performance in challenging environments with repetitive structures or low-texture surfaces, where traditional feature-based methods often struggle.

The VERF system introduces runtime monitoring techniques for pose estimation using NeRFs, enhancing pose estimation accuracy without relying on direct depth measurements. This approach demonstrates effectiveness across various scene scales, from quadruped robots to sub-orbital rockets, showcasing its potential for improving pose estimation reliability in diverse robotic applications.

NeRF-IBVS [191] combines NeRF with Image-Based Visual Servoing in a coarse-to-fine framework, achieving accurate localisation and improved performance on novel viewpoints with reduced data requirements. This approach offers significant advantages over previous techniques, enabling more robust visual localisation and navigation for autonomous robots in complex, dynamic environments.

These advancements collectively demonstrate the growing potential of NeRF-based techniques in enhancing pose estimation, localisation, and navigation capabilities for robotics, addressing key challenges in accuracy, efficiency, and adaptability to diverse environments. All the papers in the section are provided in Table 13.

Table 13
Summary of main NeRF papers used in this section.

Paper	Sensors used	Year	Platform (CPU/GPU/FPGA)
Evo-NeRF [182]	RGB Cameras (Real Setup)	2022	GPU
Dex-NeRF [183]	RGB Cameras (Real Setup)	2022	GPU
SPARTN [184]	RGB Cameras (Real Setup)	2023	GPU
GraspNeRF [185]	RGB Cameras (Real Setup)	2023	GPU
GN-Factor [186]	RGB Cameras (Simulated)	2023	GPU
CollisionNeRF [187]	RGB Cameras (Simulated)	2022	GPU
NeRF2Real [188]	RGB Cameras (Simulated)	2023	GPU
UncertaintyNeRF [189]	RGB Cameras (Simulated)	2022	GPU
GenNBV [190]	RGB Cameras (Simulated)	2024	GPU

8.2. 3D reconstruction and mapping

Recent advancements in real-time 3D reconstruction and mapping for robotics have focused on improving efficiency, scalability, and semantic understanding. VoxelCache [195] presents an innovative approach to enhance the efficiency of online mapping for robotics applications. By implementing a cache-like structure for frequently accessed voxels, the system significantly reduces computational overhead, enabling faster updates and queries of the 3D map. This method demonstrates particular promise for scenarios requiring rapid environmental understanding, such as search and rescue operations or industrial inspection tasks.

Building upon this work, OpenFusion [196] introduces an advanced framework for real-time 3D scene understanding in robotics. This research addresses the challenges of integrating Vision-Language Foundation Models (VLFMs) into robotic applications while maintaining scalability and real-time processing capabilities. By combining efficient data extraction and integration methods, OpenFusion overcomes the limitations of previous approaches, offering a more versatile and responsive system for robotic perception and decision-making in complex environments.

In scenarios with limited sensing capabilities, Li et al. [197] present a hybrid system that combines non-vision-based exploration with an active-vision-based localisation and topological mapping algorithm. This approach is particularly valuable for small robots with limited power and sensing capabilities, as demonstrated by their 70 g robot navigating 150 mm pipes. The system's ability to perform real-time localisation and mapping in such constrained environments represents a significant advancement in robotic exploration techniques for infrastructure inspection and maintenance.

Addressing the challenge of natural language understanding in spatial mapping, Huang et al. [198] introduce VLMaps, a novel spatial map representation that fuses pretrained visual-language features with 3D reconstructions of the physical environment. This method enables natural language indexing of the map without additional labelled data, allowing robots to interpret complex spatial commands. By leveraging large language models (LLMs), VLMaps can translate natural language

Table 14
Quantitative results of various NeRF 3D reconstruction approaches. Metrics include Successful Experiments percentage and F-Score. The table compares different NeRF implementations for manipulation, navigation, and general reconstruction tasks, highlighting their performance and effectiveness.

Papers	Dataset	Successful Experiments (%)	F-Score
NeRF Reconstruction for Manipulation Approaches			
Evo-NeRF [182]	self-built dataset [182]	52.6	–
Dex-NeRF [183]	self-built [183]	96.6	–
SPARTN [184]	ShapeNet [179] and YCB [176]	61.3	–
GraspNeRF [185]	self built on top of [192]	88.9	–
GN-Factor [186]	(RL-bench) simulation	44.1	–
NeRF Reconstruction for Navigation Approaches			
CollisionNeRF [187]	simulation	83.0	–
NeRF2Real [188]	simulation	78.0	–
NeRF Reconstruction Approaches			
UncertaintyNeRF [189]	NeRF Blender [12]	–	0.288
GenNBV [190]	Omniobject3d [193] and Objaverse [194]	88.0	–

instructions into sequences of open-vocabulary navigation goals and generate obstacle maps for different robot embodiments on-the-fly.

These studies collectively demonstrate the ongoing efforts to improve real-time 3D reconstruction and mapping techniques for robotics, with a focus on efficiency, scalability, and semantic understanding. As research in this field progresses, we can expect to see further integration of advanced AI techniques and more robust solutions for complex robotic navigation and interaction tasks.

8.3. Motion planning and collision

A novel approach to robot navigation in Neural Radiance Field (NeRF) environments was proposed by transforming NeRFs into equivalent Poisson Point Processes (PPPs), the authors enable rigorous quantification of uncertainty and computation of collision probabilities for robots navigating through NeRF-represented scenes. This transformation offers significant advantages in motion planning and collision mapping. The PPP representation generalises probabilistic occupancy grids to continuous volumes, aligning with the volumetric ray-tracing model underlying radiance fields. Building on this, the researchers introduce a chance-constrained trajectory optimisation method, utilising a voxel representation called the Probabilistic Unsafe Robot Region (PURR). This approach allows for fast trajectory optimisation while ensuring probabilistic safety guarantees. By combining graph-based search with spline-based trajectory optimisation, the method generates robot trajectories that satisfy user-specified collision probabilities. The real-time performance of this approach, capable of replanning at 3 Hz on a laptop, demonstrates its practical applicability in dynamic environments [199].

A transformative and groundbreaking approach to robot navigation within NeRF-represented environments was unveiled by the authors in [187]. By ingeniously adapting advanced trajectory optimisation tools to seamlessly interface with NeRF's continuous density representation, the researchers empower robots to chart and execute impeccably collision-free paths using merely RGB camera input. This cutting-edge method artfully blends the sophisticated principles of differential flatness with NeRF's density intelligence, facilitating the imposition of comprehensive pose constraints alongside dynamically feasible trajectories. The system's remarkable integration of a trajectory planner with a pose filter within an online replanning loop underscores its unparallelled robustness and extraordinary adaptability to a multitude of real-world scenarios. Diverse simulations, including those featuring quadrotor and ground robot navigation in intricate environments like jungle gyms, majestic church interiors, and the iconic Stonehenge, exemplify the remarkable versatility and formidable effectiveness of their pioneering approach.

Additionally another research also focused on Motion planning by presenting a chance-constrained trajectory optimisation method that leverages a novel transformation of Neural Radiance Fields (NeRFs) into equivalent Poisson Point Processes (PPPs). This transformation enables rigorous quantification of uncertainty in NeRFs, allowing for the computation of collision probabilities for robots navigating through NeRF-represented scenes. The method utilises a voxel representation called the Probabilistic Unsafe Robot Region (PURR) to spatially fuse chance constraints with the NeRF model, facilitating fast trajectory optimisation. By combining graph-based search with spline-based trajectory optimisation, CATNIPS generates robot trajectories that satisfy user-specified collision probabilities. The system demonstrates superior performance compared to prior works on trajectory planning in NeRF environments, with the ability to compute probabilistically safe trajectories at rates exceeding 3 Hz [199].

8.4. Video compression and motion estimation

Recent advancements in neural radiance fields (NeRFs) have demonstrated significant potential for revolutionising various applications within the industrial and robotics domains. One particularly promising application is in the realm of video compression and motion estimation.

In the context of motion estimation, the Dynamic-NeRF (D-NeRF) model [200] has shown remarkable capabilities. By training a NeRF on 3D animations of robotic arms, D-NeRF can accurately estimate depth information and motion dynamics, crucial for ensuring safe navigation in complex environments. The model's impressive performance, measured in terms of peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), highlights the potential of NeRFs to enhance a robot's situational awareness and collision avoidance capabilities.

In parallel, NeRFs have also emerged as a powerful tool for video compression. The study by Slapak et al. [201] demonstrates the effectiveness of NeRF-based techniques in achieving substantial compression savings of up to 48% for high-resolution videos. This advancement is particularly valuable in robotics, where bandwidth limitations and storage constraints are common. By reducing the size of video files, NeRF-based compression can facilitate faster data transfer, lower storage costs, and improve real-time performance.

In conclusion, NeRFs have the potential to significantly impact the field of robotics by enabling efficient video compression and accurate motion estimation. By leveraging the power of neural networks, these techniques can enhance the capabilities of robotic systems, leading to safer, more efficient, and intelligent operations.

8.5. Quantitative results

Quantitative results for various NeRF-based 3D reconstruction approaches are presented, covering tasks such as manipulation, navigation, and general reconstruction. Among the manipulation-oriented

Table 15
Summary of datasets used in the reviewed papers.

Application areas	Datasets	No. of samples	Size (GB)	Data type	Method	Best results
Human Interaction	BEHAVE [202]	321	~140 (Max)	Video	BundleSDF [157]	4.66 (CD, cm) [157]
	KIT [203]	10 264	3.9	Motion	DDGC [168]	73 (Quality of Grasp (%)) [168]
	HO3D [204]	77 558	34.3958	Image	BundleSDF [157]	0.57 (CD, cm) [157]
Object/ Shape-Based	ShapeNet [176]	51 300	71	3D Model	NodeSLAM [150], DSR-Net [167], SPARTN [184]	0.767 (IoU) [167]
	ROBI [151]	70	25.418	3D Model	ROBI [151]	~0.69 (P-to-P Err., mm) [151]
	YCB [175]	92 400	~46.431	Image	DSR-Net [167], ARM [2], S4G [173], 3DSGrasp [174], DDGC [168]	0.82 (CD, cm) [2]
	YCBInEOAT [205]	7449	22	Image	BundleSDF [157]	4.66 (CD, cm) [157]
	EGAD! [206]	2331	0.266	3D Model	DDGC [168]	0.73 (Quality of Grasp) [168]
	BUP20 [207]	~3724	35.8	Image	CoRe [166]	69.43 (F-Score, %) [166]
	Dex-Net [208]	1500	173.87	3D Model	Evo-NeRF [182], Dex-NeRF [183]	89 (Successful Grasps, %) [182]
Scene-Based	ScanNet [144]	1513	1300	3D Scan	Volumetric-Semantic-TSDF [128]	51.2 (IoU, %) [128]
	TUM [209]	88	~88.55	Video	RTSDM [129], RD3DSM [138]	89.1 (IoU, %) [129]
	KITTI360 [210]	320 000	~1200	Images	mipNeRF [211],	21.54 (PSNR, %) [210]
	EuRoC [212]	14	~12	Video	Kimera-Multi [130]	1.72 (ATE, m) [130]
	SUNRGB-D [146]	10 335	6.4	Image	Collaborative-SLAM [134]	0.78 (Mean Acc.(Max)) [134]
	Replica [213]	18	31.49	3D Model	Collaborative-SLAM [134]	0.78 (Mean Acc.(Max)) [134]
	NCLT [214]	27	~150.43	Video	S-BKI [127]	64.67 (IoU, %) [127]
	CityScapes [215]	25 000	56.819	Image	Expectation-Maximisation [135]	0.9467 (FN) [135]
	SYNTHIA [216]	220 000	98.6	Image	MR-GL [136]	7.1925 +/- 5.51 (ATE, m) [136]
	AirSim [217]	11	~12.131	3D Model	MR-GL [136]	7.1925 +/- 5.51 (ATE, m) [136]
Semantic-Based	KITTI [139]	14 999	5.27	Image	DOOR-SLAM [38], S-BKI [127], MR-GL [136]	64.67 (IoU, %) [127]
	SemanticKITTI [218]	14 999	80.189	Image	S-BKI [127], HD-CCSOM [137]	69.805 (IoU, %) [137]
	Semantic3D [140]	30	23.95	3D Scan	Automatic Scaffold 3D Reconstruction [133]	0.913 (IoU (Max), %) [133]
	KITTI360 [210]	100 000	80	3D Scan	Pointgroup [219]	58.3 (IoU (category), %) [210]
	ScanNet [144]	1513	1300	3D Scan	Volumetric-Semantic-TSDF [128]	51.2 (IoU, %) [128]
	ADE20K [220]	~27638	2.3	Image	Volumetric-Semantic-TSDF [128], Collaborative-SLAM [134]	51.2 (IoU, %) [128]
	SUNRGB-D [146]	10 335	6.4	Image	Collaborative-SLAM [134]	0.78 (Mean Acc.(Max)) [134]
	CityScapes [215]	25 000	56.819	Image	Expectation-Maximisation [135]	0.9467 (FN) [135]
	SYNTHIA [216]	220 000	98.6	Image	MR-GL [136]	7.1925 +/- 5.51 (ATE, m) [136]
	AirSim [217]	11	~12.131	3D Model	MR-GL [136]	7.1925 +/- 5.51 (ATE, m) [136]

The table categorises datasets based on their application areas, such as human interaction, object/shape-based, scene-based, and semantic-based datasets. Key metrics include the number of samples, size, data type, utilised methods, and best results achieved.YCB indicates Yale-Carnegie Mellon University (CMU)-Berkeley dataset.

methods, Dex-NeRF achieves the highest success rate (96.6%) on a self-constructed dataset, demonstrating strong robustness in object manipulation scenarios. GraspNeRF also shows competitive performance with an 88.9% success rate on a dataset derived from RL-Bench. These results are summarised in Table 14. Other methods like Evo-NeRF (52.6%) and SPARTN (61.3%) show moderate success on self-built datasets and ShapeNet/YCB datasets, respectively, while GN-Factor achieves a 44.1% success rate in RL-Bench simulations.

For navigation approaches, CollisionNeRF achieves a success rate of 83.0% in simulation, outperforming NeRF2Real, which records a 78.0% success rate. These results indicate the effectiveness of NeRF-based methods for trajectory prediction and collision avoidance in navigation tasks.

In general reconstruction approaches, UncertaintyNeRF, evaluated on the NeRF Blender dataset, achieves an F-score of 0.288, highlighting its ability to manage uncertainty during reconstruction. GenNBV demonstrates an impressive success rate of 88.0% across the OmniObject3D and Objaverse datasets, showcasing its effectiveness in capturing diverse object geometries.

Overall, NeRF techniques exhibit strong potential for manipulation, navigation, and general 3D reconstruction tasks, with notable success in datasets tailored for specific tasks. Methods like Dex-NeRF and GenNBV stand out for their high performance, while approaches like UncertaintyNeRF highlight advancements in handling uncertainty in 3D reconstructions.

9. Datasets

All datasets referenced in the proposed papers are categorised based on their contents and intended purposes. These datasets are grouped into sections according to their relevance and usage, as shown in Table 15. Most datasets are primarily compiled for categorisation algorithms,

Table 16
Quantitative results of proposed papers Part I.

Application areas	Datasets	Papers	Metric and numerical results			
Human Int	BEHAVE [202]	BundleSDF [157]	CD (cm)	AUC (%)		
			4.66	83.63		
	KIT [203]	DDGC [168]	Grasp Quality (%)	Av. Grasp Success (%)		
			73	40		
	HO3D [204]	BundleSDF [157]	CD (cm)	AUC (%)		
			0.57	96.52		
Object/Shape-Based	ShapeNet [176]	NodeSLAM [150]	Abs. Pose Error (cm)	MSE flow Err. (cm)	Acc. (mm)	IoU (%)
			1.186	–	3.484	–
		DSR-Net [167]	–	0.048	–	62.8
		SPARTN [184]	Av. Grasp Success (%)			
			61.25			
	ROBI [151]	ROBI [151]	P-to-P Err. (mm)	Scene Completeness (%)		
			~0.69	~87.65		
	YCB [175]	DSR-Net [167]	Grasp Success (%)	CD (m)	IoU (%)	
			–	–	62.8	
		ARM [2]	~45	0.017	–	
		S4G [173]	77.1	–	–	
		3DSGrasp [174]	76	0.96	–	
		DDGC [168]	40	–	–	
	YCBInEOAT [205]	BundleSDF [157]	CD (cm)	AUC (%)		
			4.66	83.63		
	EGAD! [206]	DDGC [168]	Grasp Quality (%)	Av. Grasp Success (%)		
			74	40		
	BUP20 [207]	CoRe [166]	Av. F-Score (%)	Av. Precision (%)		
			69.435	67.675		
	Dex-Net [208]	Evo-NeRF [182]	Grasp Success (%)	Trajectory Usage (%)		
			89	52.6		
		Dex-NeRF [183]	56	–		
Scene-Based	ScanNet [144]	VS-TSDF [128]	IoU (%)	Throughput (Hz)		
			51.2	53.1		
	TUM [209]	RTSDM [129]	Av. mIoU (%)	Av. Run Time (ms)		
			85.075	44.22		
		RD3DSM [138]	32.32	79.36		
	EuRoC [212]	Kimera-Multi [130]	Av. RBCD ATE (m)	Av. RBCD Time (s)		
			1.3575	224		
	SUNRGB-D [146]	Col-SLAM [134]	Colour Acc. (%)	Mean Acc. (%)		
			0.871	0.808		
	Replica [213]	Col-SLAM [134]	Colour Acc. (%)	Mean Acc. (%)		
			0.871	0.808		
	NCLT [214]	S-BKI [127]	IoU (%)	AUC (%)		
			46	78.01		
	CityScapes [215]	Exp-Max [135]	Av. Global FN	Av. Global TP		
			0.947	0.871		
	SYNTHIA [216]	MR-GL [136]	Av. ATE (m)	Av. Time (s)		
			0.216	2.620		
	AirSim [217]	MR-GL [136]	Av. ATE (m)	Av. Time (s)		
			0.216	2.620		

(*continued on next page*)

such as semantic reconstruction [127–130,133–138], and robot grasping algorithms, such as amodal reconstruction [2,166–168,173,174]. Other datasets are focused on sectors such as SLAM, Multi-View Reconstruction, and NeRF reconstruction, which utilise these methods for trajectory planning [38] and object grasping [182,183], similar to the amodal methods mentioned.

Semantic reconstruction methods utilise datasets such as KITTI [139] and ADE20K [220] due to their extensive range of semantically labelled scenes and objects, which provide suitable training for robots to classify objects within the reconstructed environment. Amodal reconstruction methods employ datasets such as the YCB [175] dataset and the ShapeNet [176] dataset to encompass a diverse array of

Table 16 (*continued*).

Application areas	Datasets	Papers	Metric and numerical results		
Semantic-Based	KITTI [139]	DOOR-SLAM [38]	Av. ATE (m)	Av. IoU (%)	Runtime (s)
			11.98	–	–
		S-BKI [127]	–	51.55	68.17
		MR-GL [136]	7.19	–	0.12
	SemanticKITTI [218]	S-BKI [127]	Av. Var. (x10^{-4})	Av. IoU (%)	AUC (%)
			–	63.4	0.78
		HD-CCSOM [137]	7.14	69.8	–
	Semantic3D [140]	AS-3D-Recons [133]	F1-Score (%)	Av. mIoU (%)	
			95.12	81.65	
	ADE20K [220]	VS-TSDF [128]	IoU (%)	Mean Acc. (%)	
			51.2	–	
		Col-SLAM [134]	–	0.808	

The table categorises the datasets based on application areas such as human interaction, object/shape-based, scene-based, and semantic-based approaches. Key metrics include Chamfer Distance (CD), Average Precision (AP), Intersection over Union (IoU), Average Running Time, and other specific performance indicators. The table highlights the effectiveness and accuracy of each method across different datasets. Human Int: Human Interaction.

objects, allowing perceived objects to be fully reconstructed to facilitate accurate grasping and avoid damage.

A brief description about each main dataset used in the paper is given below for the reader's understanding.

The KITTI dataset [139] is a comprehensive benchmark designed for autonomous driving research. It provides data collected from a moving vehicle equipped with multiple sensors, including stereo cameras, LiDAR, and GPS/IMU. The dataset covers various tasks such as depth estimation, optical flow, odometry, 3D object detection, and tracking. It consists of 22 sequences for visual odometry (11 sequences with ground truth for training and 11 without for testing) and over 200,000 annotated frames for object detection. Metrics commonly used for evaluation include absolute trajectory error (ATE), mean average precision (mAP) for object detection, and endpoint error (EPE) for optical flow.

The KITTI Semantic [147] dataset builds on the original KITTI dataset by providing pixel-level semantic annotations for a subset of scenes, enabling tasks like semantic segmentation and scene understanding. It includes over 200 annotated images with semantic labels for road, vehicles, pedestrians, and other objects relevant to urban driving. Techniques such as convolutional neural networks (CNNs) and transformer-based architectures are commonly employed for semantic segmentation tasks on this dataset. Metrics for evaluation include mean Intersection over Union (mIoU) and pixel-wise accuracy.

The EuroC dataset [145] is widely used in SLAM and visual odometry research. It consists of sequences collected from a drone equipped with stereo cameras and an inertial measurement unit (IMU). The dataset provides accurate ground truth via motion capture systems, making it suitable for benchmarking trajectory estimation techniques. Performance metrics typically include trajectory error and localisation accuracy.

The TUM-RGBD dataset [65] focuses on RGB-D data captured from Microsoft Kinect cameras, offering sequences with ground truth poses obtained via a motion capture system. It is designed for evaluating RGB-D SLAM and 3D reconstruction algorithms. Common metrics for performance evaluation include absolute trajectory error (ATE) and relative pose error (RPE).

The Replica dataset [113] is a high-quality, synthetic dataset created for scene reconstruction and robotics research. It provides photo-realistic 3D scenes rendered from RGB-D and semantic data, making it suitable for evaluating dense SLAM and reconstruction methods. The dataset supports metrics like reconstruction accuracy (e.g., Chamfer distance) and semantic segmentation accuracy.

The SUN-RGBD dataset [146] is a real-world RGB-D dataset designed for scene understanding and object detection tasks. It contains over 10,000 RGB-D images with annotated 3D bounding boxes and segmentation masks. Metrics include object detection accuracy (e.g., mAP) and semantic segmentation accuracy.

The NeRF Synthetic dataset [12] is tailored for evaluating neural radiance field (NeRF) models. It consists of synthetic 3D scenes, often with 8 to 10 objects per scene and hundreds of images per object rendered from diverse viewpoints. Train-test splits are provided to facilitate reproducibility, and common metrics include PSNR, SSIM, and LPIPS to measure rendering quality and perceptual similarity. Additionally, techniques like Gaussian Splatting are emerging for efficient representation and rendering of NeRF-like models in specific datasets.

10. Experimental results

10.1. Overview

The use of two different kitted robotic submarines [35] was employed to achieve 3D reconstruction in water using acoustic odometry and visual odometry. Tables 16 and 17 show all the quantitative results found in the proposed papers. They are organised based on the dataset they used and whether they included a dataset or not. Some papers such as [39,48,158] are not included due to the lack of quantitative data presented within these studies. On the other hand, qualitative results are shown for [39,48,158] in Tables 18, 19, 20, 21, but qualitative results for [38,184,187] are not shown due to their focus on trajectory accuracy, where quantitative results demonstrate the trajectory's precision. Below is a further explanation of each type of reconstruction.

10.2. Simultaneous localisation and mapping

Through quantitative results with other reconstruction methods on robotic systems, the SLAM method was shown to have a much-reduced runtime than other methods. With the low computation time of the SLAM process, methods like RecSLAM [40] and Collaborative SLAM [42], show the use of traditional SLAM mapping techniques and sparse reconstruction to map out the scene around the robot. Another method that has improved the use of these techniques on these systems is the use of edge computing techniques, which reduce the runtime of the algorithm, further optimising the quality of the reconstruction.

10.3. Semantic 3D reconstruction

The quantitative results for various Semantic 3D Reconstruction approaches, categorised into dense and sparse semantic reconstruction methods, are summarised in Table 7. In the category of dense semantic

Table 17
Quantitative results of proposed methods, part II.

	Datasets	Papers	Metric and numerical results of each technique	
No/Custom Database Methods	None	Collaborative-Navigation [42]	Runtime (ms)	Refresh Rate (Hz)
			~20	5
		Cloud-SLAM [41]	Power (W)	Refresh Rate (Hz)
			6	10
		RecSLAM [40]	Algorithm Runtime (ms)	Cross group edge weights
			39 (Max)	21 (Max)
		MVS-Path-Planning [152]	Av. F-Score (%)	Runtime (s)
			84.48	4.66
		RGBD-SLAM-Agriculture [34]	Mean Euclidean Distance (m)	
			2.473	
		Reconstruction in Welding [153]	Error Rate (%)	Runtime (s)
			9.03	5.26
		kPAM-SC [169]	Av. False Negative Rate (%)	Av. Failure Rate (%)
			2	26
		Uncertain Object Ins-Segm [172]	Av. SE (cm)	Grasp Success Rate (%)
			0.508	94.7
		NeRF-SLAM [187]	Failure Rate (%)	Rotational Err. (%)
			~15	~5
		GNFactor [186]	Success Rate (%)	Multi-Task Success Rate (%)
			33.3	54.8
	Custom	SLAM-Box [33]	Runtime (ms)	Power (W)
			38	13.8
		Underwater-Visual-SLAM [35]	Av. Root Means Squared Error (RMSE) (m)	Av. Refresh Rate (Hz)
			0.27	5.3 and 20.8 (2)
		Topomap [47]	Runtime (s)	Refresh Rate (Hz)
			86.1 (Max)	20–30
		3D-OOI-SLAM [13]	Mean Elevation ang. Error (deg)	Mean stand off dist. error (m)
			3.65	0.05
		Plan3D [148]	Av. F-Score (%)	Runtime (s)
			79.87	4800 (Max)
		Sequential-Reconstruciton [149]	Av. Accuracy (m)	
			0.03484	
		Shape-Completion [170]	Av. IoU (%)	Av. Centre Dev. (mm)
			61	5.7
		NeRF2Real [188]	Success Rate (%)	Localisation Err. (m)
			87	0.27
		GraspNeRF [185]	Success Rate (%)	
			88.9	
	Dataset from [12]	Uncertain-NeRF [189]	Av. F-Score (%)	
			28.8	

Key metrics include runtime, refresh rate, power consumption, error rates, and success rates. The table compares the performance of various approaches, highlighting their effectiveness and efficiency in different scenarios.

reconstruction, RA-SLAM [173] achieves an Intersection over Union (IoU) of 51.2%, while S-BKI [127] and HD-CCSOM [137] report IoU values of 64.67% and 62.5%, respectively. RTSDM [129] stands out with an IoU of 85.1% and a mean Absolute Trajectory Error (ATE) of 0.11345 metres, demonstrating high accuracy in dense reconstruction. Additionally, CrossSemantic3D [134] achieves a mean accuracy of 73.8%, and RD3DSM [138] reports an IoU of 32.32% with a maximum collection runtime of 47.83 ms.

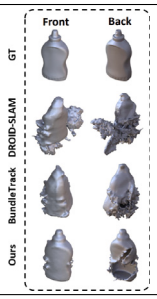
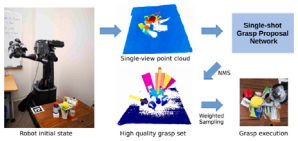
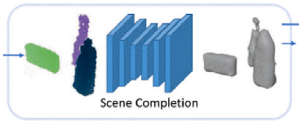
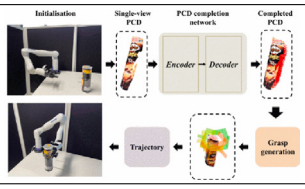
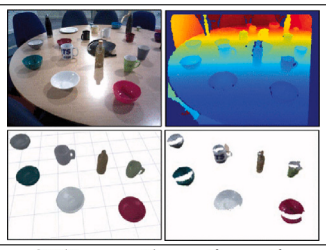
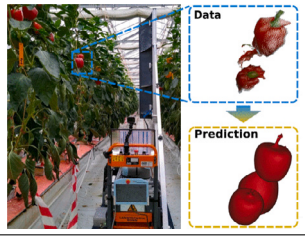
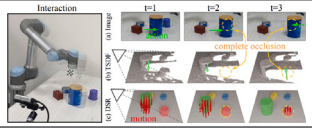
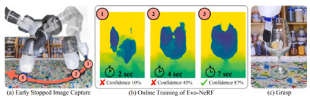
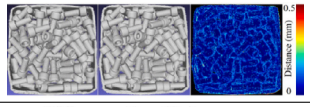
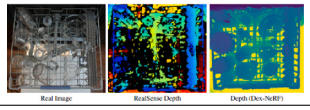
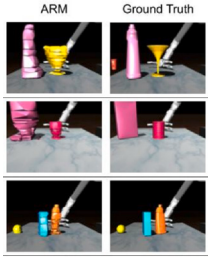
In the realm of sparse semantic reconstruction, Kimera-Multi [130] achieves a maximum mean accuracy of 89.53% with a collection runtime of 3125 ms using the RBCD (ES) method, and a mean ATE of 1.54025 metres. The 3DSemanticRobotDog [133] demonstrates an IoU of 84%, while MR-SLAM shows a collection runtime of 248 ms for car and UAV and a mean ATE of 2.12 metres. These results indicate that sparse semantic reconstruction methods, while generally quicker, may vary in accuracy depending on the approach and application.

10.4. Multi-view 3D reconstruction

The quantitative results for various Multi-View 3D Reconstruction approaches are summarised in Table 9. These methods leverage multiple viewpoints to achieve high accuracy in reconstructing 3D models. Plan3D
[148] achieves an F-Score of 79.22%, demonstrating its effectiveness in creating detailed 3D models.

NodeSLAM [150] achieves an average accuracy of 0.03484 metres, highlighting its precision in reconstruction tasks. ROBI [158] reports a distance error of 0.5775 metres using a zigzag pattern, indicating

Table 18
Qualitative results in proposed papers Part I.

Paper	Image	Paper	Image
BundleSDF [157]		S4G [173]	
DDGC [168]		3DSGrasp [174]	
NodeSLAM [150]		CoRe [166]	
DSR-Net [167]		Evo-NeRF [182]	
ROBI [151]		Dex-NeRF [183]	
ARM [2]		VS-TSDF [128]	

This table includes the qualitative results of various proposed methods. BundleSDF [157] employs a signed distance function for bundle adjustment. S4G [173] utilises semantic scene graph generation. DDGC [168] focuses on dense geometric correspondence. 3DSGrasp [174] integrates 3D semantic grasping. NodeSLAM [150] combines SLAM with semantic mapping. CoRe [166] uses contrastive learning for reconstruction. DSR-Net [167] applies depth-supervised reconstruction networks. Evo-NeRF [182] involves evolutionary neural radiance fields. ROBI [151] targets robust object manipulation. Dex-NeRF [183] focuses on dexterous manipulation using NeRF. ARM [2] employs amodal reconstruction methods. VS-TSDF [128] utilises voxel-based signed distance functions for real-time reconstruction. Each image represents the qualitative outcomes of these techniques.

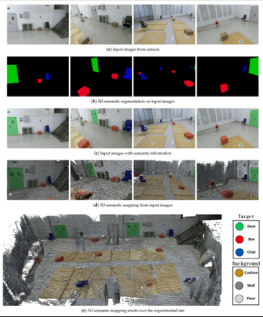
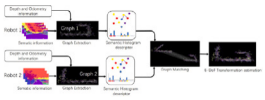
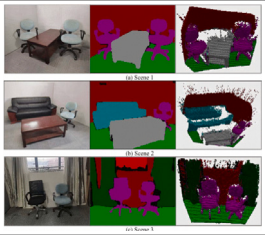
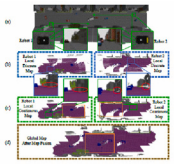
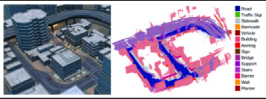
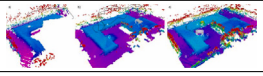
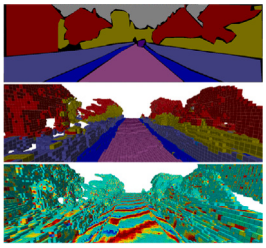
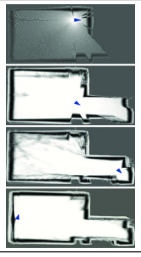
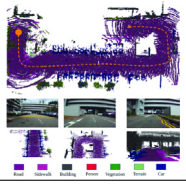
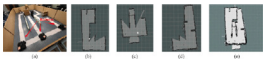
its capability to handle complex object geometries. Additionally, [149] achieves an average accuracy of 0.1867 metres, while [152,153] report F-Scores of 84.48% and error rates of 9.03%, respectively. These results underscore the varying levels of accuracy and error rates among different Multi-View 3D Reconstruction approaches.

10.5. Amodal 3D reconstruction

The quantitative results for various Amodal 3D Reconstruction approaches, categorised into Voxel/Mesh Shape Completed Reconstruction Approaches and Point Cloud Shape Completed Reconstruction Approaches, are summarised in Table 12. In the Voxel/Mesh category, DSR-Net [167] achieves an Intersection over Union (IoU) of 62% with a Mean Squared Error (MSE) of 0.048 cm, indicating high accuracy in shape completion tasks. ARM [2] and DDGC [168] report success grasp rates of 42% and 40%, respectively, although they do not provide IoU or error metrics.

In the Point Cloud Shape category, kPAM-SC [169] stands out with a success grasp rate of 96%, demonstrating significant improvement in grasp success rates through point cloud shape completion. Similarly, the method presented in [170] achieves an IoU of 61%, while [172] reports a success grasp rate of 94.7% and an MSE of 0.508 cm. S4G [173] and 3DSGrasp [170] showcase their effectiveness in enhancing the robot's ability to accurately manipulate objects with success grasp rates of 92.5% and 76%, respectively. These results highlight the varying

Table 19
Qualitative results in proposed papers Part II.

Paper	Image	Paper	Image
RTSDM [129]		MR-GL [136]	
RD3DSM [138]		HD-CCSOM [137]	
Kimera-Multi [130]		AS-3D-Recons [133]	
Col-SLAM [134]		Collaborative-Navigation [42]	
S-BKI [127]		Cloud-SLAM [41]	
Exp-Max [135]		RecSLAM [40]	

This table includes the qualitative results of various proposed methods. RTSDM [129] focuses on real-time semantic depth maps. MR-GL [136] utilises multi-resolution geometric learning. RD3DSM [138] employs real-time dense 3D semantic mapping. HD-CCSOM [137] focuses on high-definition cognitive consistent semantic object mapping. Kimera-Multi [130] integrates multi-robot systems for 3D reconstruction. AS-3D-Recons [133] uses deep learning for 3D reconstruction. Col-SLAM [134] combines collaborative SLAM with semantic mapping. Collaborative-Navigation [42] explores multi-robot collaborative navigation. S-BKI [127] utilises Bayesian kernel inference for semantic mapping. Cloud-SLAM [41] offloads SLAM computations to the cloud. Exp-Max [135] employs hierarchical expectation maximisation for semantic mapping. RecSLAM [40] focuses on edge computing for SLAM. Each image represents the qualitative outcomes of these techniques.

levels of accuracy and success rates among different Amodal 3D Reconstruction approaches, reflecting their effectiveness in shape completion and object manipulation tasks.

10.6. NeRF 3D reconstruction

The quantitative results for various NeRF 3D Reconstruction approaches, divided into categories for manipulation and navigation tasks, are summarised in Table 14.

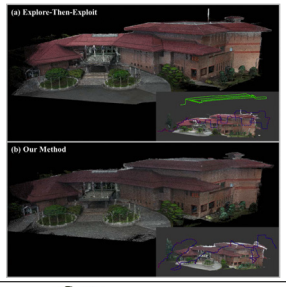
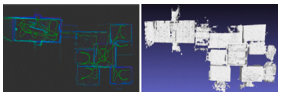
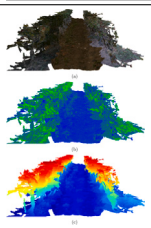
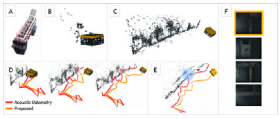
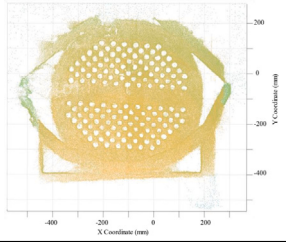
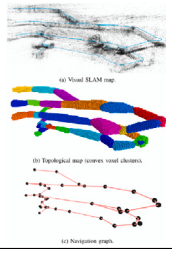
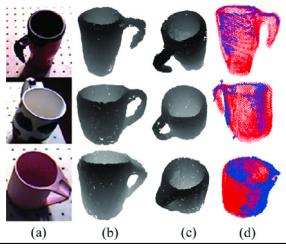
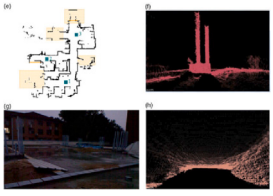
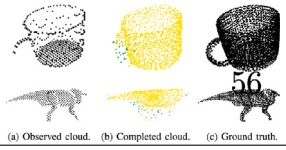
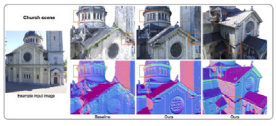
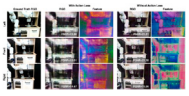
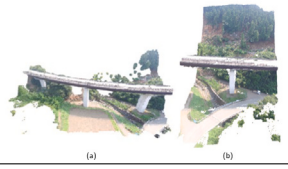
In the NeRF Reconstruction for Manipulation Approaches category, Evo-NeRF [182] achieves a successful experiment rate of 52.6%, while Dex-NeRF [183] significantly outperforms with a success rate of 96.6%. SPARTN [184] and GraspNeRF [185] report success rates of 61.3% and 88.9%, respectively, highlighting their effectiveness in object manipulation tasks. GN-Factor [186] achieves a success rate of 44.1%, indicating room for improvement.

For NeRF Reconstruction for Navigation Approaches, Collision-NeRF [187] and NeRF2Real [188] demonstrate high success rates of 83.0% and 78.0%, respectively, showcasing their reliability in navigation tasks. UncertaintyNeRF [189] is noted for its unique approach, achieving an F-Score of 0.288, though more exploration is suggested in this area. These results highlight the varying levels of effectiveness in NeRF 3D Reconstruction approaches for both manipulation and navigation tasks.

11. Discussion

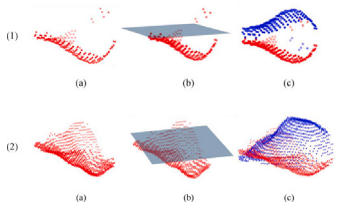
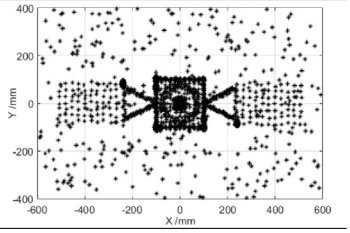
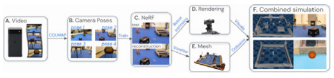
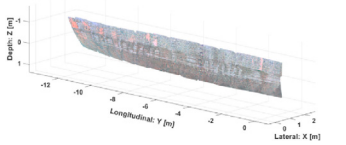
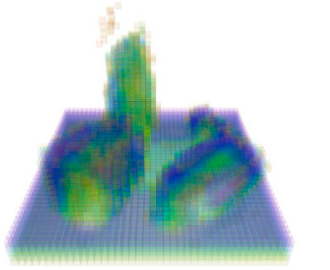
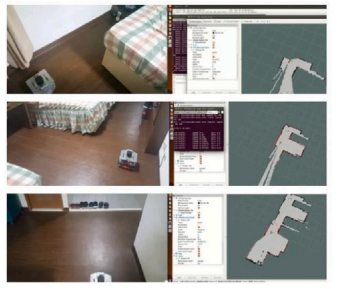
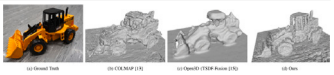
The field of 3D Semantic Reconstruction has markedly progressed, offering significant enhancements in how robots perceive and interact with complex environments. Innovations such as SLAM (Simultaneous Localisation and Mapping) and NeRF (Neural Radiance Fields) have

Table 20
Qualitative results in proposed papers Part III.

Paper	Image	Paper	Image
MVS-PAth-Planning [152]		SLAM-Box [33]	
RGBD-SLAM-Agriculture [34]		UW-Vis-SLAM [35]	
Reconstruction in Welding [153]		Topomap [47]	
kPAM-SC [169]		3D-Vis-Map [48]	
Unc-Obj-Ins-Seg [172]		Plan3D [148]	
GNFactor [186]		Seq-Rec [149]	

This table includes the qualitative results of various proposed methods. MVS-Path-Planning [152] employs multi-view stereo path planning. SLAM-Box [33] focuses on energy-efficient SLAM in a box framework. RGBD-SLAM-Agriculture [34] integrates RGB-D SLAM for agricultural applications. UW-Vis-SLAM [35] utilises underwater visual SLAM. Reconstruction in Welding [153] focuses on novel reconstruction techniques for welding applications. Topomap [47] employs topographic mapping for SLAM. kPAM-SC [169] uses keypoint affordances for manipulation. 3D-Vis-Map [48] integrates visual mapping in 3D. Unc-Obj-Ins-Seg [172] focuses on uncertain object instance segmentation. Plan3D [148] employs planning in 3D environments. GNFactor [186] utilises graph neural factors for SLAM. Seq-Rec [149] focuses on sequential reconstruction methods.

Table 21
Qualitative results in proposed papers Part IV.

Paper	Image	Paper	Image
Shp-Comp [170]		Dual-Arm [158]	
NeRF2Real [188]		3D-OOI-SLAM [13]	
GraspNeRF [185]		Hybrid-SLAM [39]	
Uncert-NeRF [189]			

This table includes the qualitative results of various proposed methods. Shp-Comp [170] employs symmetry-completion techniques. Dual-Arm [158] focuses on virtual dual-arm manipulation. NeRF2Real [188] utilises NeRF for real-world applications. 3D-OOI-SLAM [13] integrates object-oriented SLAM. GraspNeRF [185] focuses on grasping using NeRF. Hybrid-SLAM [39] employs hybrid SLAM techniques. Each image represents the qualitative outcomes of these techniques.

been instrumental, merging depth mapping with semantic understanding to elevate machine autonomy. Despite these advances, the area faces substantial challenges, notably in computational intensity and data integration. This complexity often impedes real-time processing, particularly critical in dynamic and unpredictable settings. Moreover, the availability of diverse, comprehensive datasets remains a significant limitation. Most current datasets do not adequately represent the variety of scenarios that robots might encounter, restricting the development and performance of reconstruction systems.

Particularly, robotic applications in disaster response and healthcare could benefit from focused advancements in 3D reconstruction technologies to meet the high demands for reliability and precision. These applications require systems that can operate flawlessly under stressful and variable conditions. Future research could benefit from integrating AI to enhance semantic understanding and decision-making capabilities. Developing algorithms that optimise computational efficiency while boosting real-time processing capabilities will be essential. Additionally, exploring unsupervised learning methods could pave the way for robots to adapt to new environments more efficiently, without heavy reliance on extensive labelled datasets. Addressing these challenges and opportunities could significantly expand the scope and effectiveness of robotic applications, making them more versatile and capable across various industries.

In addition to the aforementioned challenges, another critical limitation discussed in the paper is the difficulty in handling occlusions and partial views during 3D reconstruction. Current methods often struggle to accurately reconstruct objects that are partially obscured, which is common in real-world applications. This limitation significantly affects the reliability of the generated models, particularly in cluttered and dynamic environments. Moreover, there is a notable lack of integration between 3D reconstruction algorithms and advanced sensor technologies, such as thermal imaging, which could provide additional layers of information crucial for certain applications.

Future research should prioritise developing algorithms capable of predicting and filling in missing data from occluded or partially visible objects using contextual information. Additionally, integrating thermal and multispectral imaging with traditional RGB and depth sensors could enhance the robustness and versatility of 3D reconstruction systems. Addressing these limitations will be pivotal in advancing the field and expanding the practical applications of 3D reconstruction in robotics.

12. Conclusion

This paper aims to present a comprehensive review of the latest advancements in 3D reconstruction within robotics, highlighting its critical importance in enhancing robotic perception and interaction with complex environments. By synthesising insights from various research works, this review seeks to demonstrate how cutting-edge techniques like SLAM and NeRF contribute to overcoming existing challenges, such as computational demands and data integration, while addressing the need for diverse datasets. The collective analysis of these papers underscores the significance of innovative algorithms and unsupervised learning methods, paving the way for future progress in the field.

The exploration of 3D reconstruction in robotics reveals a dynamic interplay of advancements and challenges. While developments like SLAM and NeRF have significantly propelled robotic capabilities in understanding and interacting with complex environments, hurdles in computational demands and data integration persist. The lack of diverse, realistic datasets further complicates the advancement of this

technology. Addressing these issues, alongside exploring innovative algorithms and unsupervised learning methods, will be crucial for future progress. Ultimately, the continued evolution in 3D reconstruction promises to enhance the autonomy and effectiveness of robots across various industries, making this an exciting field of ongoing research and application.

CRediT authorship contribution statement

Dharmendra Selvaratnam: Writing – review & editing, Writing – original draft, Conceptualization. **Dena Bazazian:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the PhD studentship funding supporting Dharmendra Selvaratnam's research, provided through the EPSRC Doctoral Training Partnership (DTP) via the School of Engineering, Computing and Mathematics (SECaM), University of Plymouth. Furthermore, we thank Christopher Jenner for his initial contributions to this paper.

Data availability

No data was used for the research described in the article.

References

[1] Ann NQ, Achmad MH, Bayuaji L, Daud MR, Pebrianti D. Study on 3D scene reconstruction in robot navigation using stereo vision. In: 2016 IEEE international conference on automatic control and intelligent systems. IEEE; 2016, p. 72–7.
[2] Agnew W, Xie C, Walsman A, Murad O, Wang Y, Domingos P, et al. Amodal 3D reconstruction for robotic manipulation via stability and connectivity. In: Conference on robot learning. PMLR; 2021, p. 1498–508.
[3] Wang Y, James S, Stathopoulou EK, Beltrán-González C, Konishi Y, Del Bue A. Autonomous 3-d reconstruction, mapping, and exploration of indoor environments with a robotic arm. IEEE Robot Autom Lett 2019;4(4):3340–7.
[4] Yandun F, Silwal A, Kantor G. Visual 3d reconstruction and dynamic simulation of fruit trees for robotic manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020, p. 54–5.
[5] Zhao C, Sun L, Stolkin R. A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition. In: 2017 18th international conference on advanced robotics. IEEE; 2017, p. 75–82.
[6] Ashokaraj I, Tsourdos A, Silson P, White B. Sensor based robot localisation and navigation: Using interval analysis and extended Kalman filter. In: 2004 5th Asian control conference (IEEE cat. no. 04EX904), vol. 2, IEEE; 2004, p. 1086–93.
[7] Schmied A, Fischer T, Danelljan M, Pollefeys M, Yu F. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 3216–26.
[8] Sun Y, Huang Z, Zhang H, Cao Z, Xu D. 3DRIMR: 3D reconstruction and imaging via mmWave radar based on deep learning. In: 2021 IEEE international performance, computing, and communications conference. IEEE; 2021, p. 1–8.
[9] Zhang Y, Tosi F, Mattoccia S, Poggi M. Go-slam: Global optimization for consistent 3d instant reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 3727–37.
[10] Xiong W, Yang H, Zhou P, Fu K, Zhu J. Spatiotemporal correlation-based accurate 3D face imaging using speckle projection and real-time improvement. Appl Sci 2021;11(18):8588.
[11] Pritsker AAB. Introduction to simulation and SLAM II. Halsted Press; 1984.
[12] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. Commun ACM 2021;65(1):99–106.
[13] Wang H, Zhang C, Song Y, Pang B, Zhang G. Three-dimensional reconstruction based on visual SLAM of mobile robot in search and rescue disaster scenarios. Robotica 2020;38(2):350–73.
[14] Matsuki H, Murai R, Kelly PH, Davison AJ. Gaussian splatting slam. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 18039–48.
[15] Picard Q, Chevobbe S, Darouich M, Didier J-Y. A survey on real-time 3D scene reconstruction with SLAM methods in embedded systems. 2023, arXiv preprint arXiv:2309.05349.
[16] Samavati T, Soryani M. Deep learning-based 3D reconstruction: a survey. Artif Intell Rev 2023;56(9):9175–219.
[17] Macario Barros A, Michel M, Moline Y, Corre G, Carrel F. A comprehensive survey of visual slam algorithms. Robotics 2022;11(1):24.
[18] Cai D, Li R, Hu Z, Lu J, Li S, Zhao Y. A comprehensive overview of core modules in visual SLAM framework. Neurocomputing 2024;127760.
[19] Zhang Y, Shi P, Li J. 3D LiDAR SLAM: A survey. Photogramm Rec 2024.
[20] Saeedi S, Bodin B, Wagstaff H, Nisbet A, Nardi L, Mawer J, et al. Navigating the landscape for real-time localization and mapping for robotics and virtual and augmented reality. Proc IEEE 2018;106(11):2020–39.
[21] Maboudi M, Homaei M, Song S, Malihi S, Saadatseresht M, Gerke M. A review on viewpoints and path planning for UAV-based 3D reconstruction. IEEE J Sel Top Appl Earth Obs Remote Sens 2023.
[22] Liu S, Liu L, Tang J, Yu B, Wang Y, Shi W. Edge computing for autonomous driving: Opportunities and challenges. Proc IEEE 2019;107(8):1697–716.
[23] Zaffar M, Ehsan S, Stolkin R, Maier KM. Sensors, SLAM and long-term autonomy: A review. In: 2018 NASA/ESA conference on adaptive hardware and systems. IEEE; 2018, p. 285–90.
[24] Kang Z, Yang J, Yang Z, Cheng S. A review of techniques for 3D reconstruction of indoor environments. ISPRS Int J Geo-Inf 2020;9(5):330.
[25] Cui Y, Chen R, Chu W, Chen L, Tian D, Li Y, et al. Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE Trans Intell Transp Syst 2021;23(2):722–39.
[26] Mur-Artal R, Tardós JD. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Trans Robot 2017;33(5):1255–62.
[27] Campos C, Elvira R, Rodríguez JJG, Montiel JM, Tardós JD. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Trans Robot 2021;37(6):1874–90.
[28] Sumikura S, Shibuya M, Sakurada K. OpenVSLAM: A versatile visual SLAM framework. In: Proceedings of the 27th ACM international conference on multimedia. 2019, p. 2292–5.
[29] Labbé M, Michaud F. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. J Field Robot 2019;36(2):416–46.
[30] Xu K, Hao Y, Yuan S, Wang C, Xie L. Airslam: An efficient and illumination-robust point-line visual slam system. 2024, arXiv preprint arXiv:2408.03520.
[31] Lim H, Jeon J, Myung H. UV-SLAM: Unconstrained line-based SLAM using vanishing points for structural mapping. IEEE Robot Autom Lett 2022;7(2):1518–25.
[32] Fu T, Su S, Lu Y, Wang C. Islam: Imperative slam. IEEE Robot Autom Lett 2024.
[33] Flottmann M, Eisoldt M, Gaal J, Rothmann M, Tassemeier M, Wiemann T, et al. Energy-efficient FPGA-accelerated LiDAR-based SLAM for embedded robotics. In: 2021 international conference on field-programmable technology. IEEE; 2021, p. 1–6.
[34] Vulpi F, Marani R, Petitti A, Reina G, Milella A. An RGB-D multi-view perspective for autonomous agricultural robots. Comput Electron Agric 2022;202:107419.
[35] Vargas E, Scona R, Willners JS, Luczynski T, Cao Y, Wang S, et al. Robust underwater visual SLAM fusing acoustic sensing. In: 2021 IEEE international conference on robotics and automation. IEEE; 2021, p. 2140–6.
[36] Wang Z, Cheng Q, Mu X. RU-SLAM: A robust deep-learning visual simultaneous localization and mapping (SLAM) system for weakly textured underwater environments. Sensors 2024;24(6):1937.
[37] Yang K, Zhang Z, Cui R, Yan W. Acoustic-optic assisted multisensor navigation for autonomous underwater vehicles. Ocean Eng 2024;297:117139.
[38] Lajoie P-Y, Ramtoula B, Chang Y, Carlone L, Beltrame G. DOOR-SLAM: Distributed, online, and outlier resilient SLAM for robotic teams. IEEE Robot Autom Lett 2020;5(2):1656–63.
[39] Mac TT, Lin C-Y, Huan NG, Duc L, Nhat PCH, Hai HH. Hybrid SLAM-based exploration of a mobile robot for 3D scenario reconstruction and autonomous navigation. Acta Polytech Hung 2021;18:197–212.
[40] Huang P, Zeng L, Chen X, Luo K, Zhou Z, Yu S. Edge robotics: Edge-computing-accelerated multirobot simultaneous localization and mapping (SLAM). IEEE Internet Things J 2022;9(15):14087–102.
[41] Sarker VK, Queralta JP, Gia TN, Tenhunen H, Westerlund T. Offloading SLAM for indoor mobile robots with edge-fog-cloud computing. In: 2019 1st international conference on advances in science, engineering and robotics technology. IEEE; 2019, p. 1–6.

[42] Fankhauser P, Bloesch M, Krüsi P, Diethelm R, Wermelinger M, Schneider T, et al. Collaborative navigation for flying and walking robots. In: 2016 IEEE/RSJ international conference on intelligent robots and systems. IEEE; 2016, p. 2859–66.

[43] Ress V, Zhang W, Skuddis D, Haala N, Soergel U. SLAM for indoor mapping of wide area construction environments. 2024, arXiv preprint arXiv:2404.17215.

[44] Lu Q, Pan Y, Hu L, He J. A method for reconstructing background from RGB-D SLAM in indoor dynamic environments. Sensors 2023;23(7):3529.

[45] Huang Y, Li P, Ma S, Yan S, Tan M, Yu J, et al. Visual-inertial-acoustic sensor fusion for accurate autonomous localization of underwater vehicles. IEEE Trans Cybern 2024.

[46] Zhang J, Han F, Han D, Yang J, Zhao W, Li H. Integration of sonar and visual inertial systems for SLAM in underwater environments. IEEE Sens J 2024.

[47] Blochliger F, Fehr M, Dymczyk M, Schneider T, Siegwart R. Topomap: Topological mapping and navigation based on visual SLAM maps. In: 2018 IEEE international conference on robotics and automation. IEEE; 2018, p. 3818–25.

[48] Hong S, Kim J. Three-dimensional visual mapping of underwater ship hull surface using piecewise-planar SLAM. Int J Control Autom Syst 2020;18:564–74.

[49] Liu P, Geppert M, Heng L, Sattler T, Geiger A, Pollefeys M. Towards robust visual odometry with a multi-camera system. In: 2018 IEEE/RSJ international conference on intelligent robots and systems. IEEE; 2018, p. 1154–61.

[50] Wang X, Wang C, Liu B, Zhou X, Zhang L, Zheng J, et al. Multi-view stereo in the deep learning era: A comprehensive review. Displays 2021;70:102102.

[51] Won C, Seok H, Cui Z, Pollefeys M, Lim J. OmniSLAM: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems. In: 2020 IEEE international conference on robotics and automation. IEEE; 2020, p. 559–66.

[52] Chen X, Lu H, Xiao J, Zhang H. Distributed monocular multi-robot slam. In: 2018 IEEE 8th annual international conference on CYBER technology in automation, control, and intelligent systems. IEEE; 2018, p. 73–8.

[53] Dong Y, Ding H, Zha F, Li M. A practical multi-camera SLAM system for large mobile robots. In: 2022 2nd international conference on big data, artificial intelligence and risk management. IEEE; 2022, p. 179–84.

[54] Yang AJ, Cui C, Bârsan IA, Urtasun R, Wang S. Asynchronous multi-view SLAM. In: 2021 IEEE international conference on robotics and automation. IEEE; 2021, p. 5669–76.

[55] Song H, Liu C, Dai H. Bundledslam: An accurate visual slam system using multiple cameras. In: 2024 IEEE 7th advanced information technology, electronic and automation control conference, vol. 7, IEEE; 2024, p. 106–11.

[56] Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, et al. The EuRoC micro aerial vehicle datasets. Int J Robot Res 2016;35(10):1157–63.

[57] Li S, Pang L, Hu X. Multicam-SLAM: Non-overlapping multi-camera SLAM for indirect visual localization and navigation. 2024, arXiv preprint arXiv:2406.06374.

[58] Yang Y, Tang D, Wang D, Song W, Wang J, Fu M. Multi-camera visual SLAM for off-road navigation. Robot Auton Syst 2020;128:103505.

[59] Hebert MH, Thorpe CE, Stentz A. Intelligent unmanned ground vehicles: autonomous navigation research at Carnegie Mellon, vol. 388, Springer Science & Business Media; 2012.

[60] Keetha N, Karhade J, Jatavallabhula KM, Yang G, Scherer S, Ramanan D, et al. SplaTAM: Splat track & map 3D Gaussians for dense RGB-D SLAM. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 21357–66.

[61] Yan C, Qu D, Xu D, Zhao B, Wang Z, Wang D, et al. Gs-slam: Dense visual slam with 3d gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 19595–604.

[62] Wu T, Yuan Y-J, Zhang L-X, Yang J, Cao Y-P, Yan L-Q, et al. Recent advances in 3d gaussian splatting. Comput Vis Media 2024;10(4):613–42.

[63] Straub J, Whelan T, Ma L, Chen Y, Wijmans E, Green S, et al. The replica dataset: A digital replica of indoor spaces. 2019, arXiv preprint arXiv:1906.05797.

[64] Kasar A. Benchmarking and comparing popular visual SLAM algorithms. 2018, arXiv preprint arXiv:1811.09895.

[65] Sun S, Mielle M, Lilienthal AJ, Magnusson M. High-fidelity SLAM using Gaussian splatting with rendering-guided densification and regularized optimization. 2024, arXiv preprint arXiv:2403.12535.

[66] Jiang W, Lei B, Ashton K, Daniilidis K. AG-SLAM: Active Gaussian splatting SLAM. 2024, arXiv preprint arXiv:2410.17422.

[67] Ahmed MF, Masood K, Fremont V, Fantoni I. Active slam: A review on last decade. Sensors 2023;23(19):8097.

[68] Zhu Z, Peng S, Larsson V, Xu W, Bao H, Cui Z, et al. Nice-slam: Neural implicit scalable encoding for slam. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 12786–96.

[69] Cartillier V, Schindler G, Essa I. SLAIM: Robust dense neural SLAM for online tracking and mapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 2862–71.

[70] Jiang H, Xu Y, Li K, Feng J, Zhang L. Rodyn-slam: Robust dynamic dense rgb-d slam with neural radiance fields. IEEE Robot Autom Lett 2024.

[71] Ruan C, Zang Q, Zhang K, Huang K. Dn-slam: A visual slam with orb features and nerf mapping in dynamic environments. IEEE Sens J 2023.

[72] Vial P, Palomeras N, Solà J, Carreras M. Underwater Pose SLAM using GMM scan matching for a mechanical profiling sonar. J Field Robot 2024;41(3):511–38.

[73] Zhuang L, Chen X, Lu W, Yan Y. Graph matching for underwater simultaneous localization and mapping using multibeam sonar imaging. J Mar Sci Eng 2024;12(10):1859.

[74] ASL. GitHub library for ASL's VI sensor. 2018, URL https://github.com/ethz-asl/libvisensor/.

[75] Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans Robot 2015;31(5):1147–63.

[76] Rosinol A, Violette A, Abate M, Hughes N, Chang Y, Shi J, et al. Kimera: From SLAM to spatial perception with 3D dynamic scene graphs. Int J Robot Res 2021;40(12–14):1510–46.

[77] Pumarola A, Vakhitov A, Agudo A, Sanfeliu A, Moreno-Noguer F. PL-SLAM: Real-time monocular visual SLAM with points and lines. In: 2017 IEEE international conference on robotics and automation. IEEE; 2017, p. 4503–8.

[78] Engel J, Koltun V, Cremers D. Direct sparse odometry. IEEE Trans Pattern Anal Mach Intell 2017;40(3):611–25.

[79] Forster C, Zhang Z, Gassner M, Werlberger M, Scaramuzza D. SVO: Semidirect visual odometry for monocular and multicamera systems. IEEE Trans Robot 2016;33(2):249–65.

[80] Zubizarreta J, Aguinaga I, Montiel JMM. Direct sparse mapping. IEEE Trans Robot 2020;36(4):1363–70.

[81] Lipson L, Teed Z, Deng J. Deep patch visual slam. In: European conference on computer vision. Springer; 2025, p. 424–40.

[82] Zhou H, Ummenhofer B, Brox T. DeepTAM: Deep tracking and mapping with convolutional neural networks. Int J Comput Vis 2020;128(3):756–69.

[83] Wang W, Hu Y, Scherer S. Tartanvo: A generalizable learning-based vo. In: Conference on robot learning. PMLR; 2021, p. 1761–72.

[84] Teed Z, Deng J. Deepv2d: Video to depth with differentiable structure from motion. 2018, arXiv preprint arXiv:1812.04605.

[85] Czarnowski J, Laidlow T, Clark R, Davison AJ. Deepfactors: Real-time probabilistic dense monocular slam. IEEE Robot Autom Lett 2020;5(2):721–8.

[86] Huang H, Li L, Cheng H, Yeung S-K. Photo-SLAM: Real-time simultaneous localization and photorealistic mapping for monocular stereo and RGB-D cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 21584–93.

[87] Sandström E, Li Y, Van Gool L, Oswald MR. Point-slam: Dense neural point cloud-based slam. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 18433–44.

[88] Ha S, Yeon J, Yu H. Rgbd gs-icp slam. In: European conference on computer vision. Springer; 2025, p. 180–97.

[89] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: 2007 6th IEEE and ACM international symposium on mixed and augmented reality. IEEE; 2007, p. 225–34.

[90] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM. In: European conference on computer vision. Springer; 2014, p. 834–49.

[91] Endres F, Hess J, Engelhard N, Sturm J, Cremers D, Burgard W. An evaluation of the RGB-D SLAM system. In: 2012 IEEE international conference on robotics and automation. IEEE; 2012, p. 1691–6.

[92] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. ACM Trans Graph 2023;42(4):1–139.

[93] Chen T, Shorinwa O, Zeng W, Bruno J, Dames P, Schwager M. Splat-nav: Safe real-time robot navigation in Gaussian splatting maps. 2024, arXiv preprint arXiv:2403.02751.

[94] Cap M, Novak P, Kleiner A, Selecky M. Prioritized planning algorithms for trajectory coordination of multiple mobile robots. IEEE Trans Autom Sci Eng 2015;12(3):835–49.

[95] Chen N, Wang M, Alkim T, Van Arem B. Formation control for connected and automated vehicles on multi-lane roads: Relative motion planning and conflict resolution. IEEE Trans Intell Transp Syst 2020.

[96] Wu Y, Wu Y, Gkioxari G, Tian Y, Tamar A, Russell S. Learning a semantic prior for guided navigation. In: European conference on computer vision. 2018.

[97] Zhang J, Wei Z, Fan J, Peng J. Curriculum learning for vision-and-language navigation. In: Advances in neural information processing systems. 2021.

[98] Li Y, Pathak D. Object-aware Gaussian splatting for robotic manipulation. In: ICRA 2024 workshop on 3D visual representations for robot manipulation.

[99] Mitash C, Hussein M, Vanbaar J, Terhuja V, Katyal K. Scaling object-centric robotic manipulation with multimodal object identification. 2024.

[100] Abou-Chakra J, Rana K, Dayoub F, Sünderhauf N. Physically embodied Gaussian splatting: A realtime correctable world model for robotics. 2024, arXiv preprint arXiv:2406.10788.

[101] Zhang C, Zhou Y, Zhang L. Voxel-mesh hybrid representation for real-time view synthesis by meshing density field. IEEE Trans Vis Comput Graphics 2024;(01):1–13.

[102] Mello Rella E, Chhatkuli A, Konukoglu E, Van Gool L. Neural vector fields for implicit surface representation and inference. Int J Comput Vis 2024;1–24.

[103] Lu G, Zhang S, Wang Z, Liu C, Lu J, Tang Y. ManiGaussian: Dynamic Gaussian splatting for multi-task robotic manipulation. 2024, arXiv preprint arXiv:2403.08321.

[104] Lin J. Dynamic NeRF: A review. 2024, arXiv preprint arXiv:2405.08609.
[105] Kazerouni IA, Fitzgerald L, Dooly G, Toal D. A survey of state-of-the-art on visual SLAM. Expert Syst Appl 2022;205:117734.
[106] Matsuki H, Murai R, Kelly PHJ, Davison AJ. Gaussian splatting SLAM. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.
[107] Bowser S, Lukin SM. 3D Gaussian splatting for human-robot interaction. In: The 1st interAI workshop: interactive AI for human-centered robotics. 2024.
[108] Zhan K, Lang X, Zhou X, et al. Street Gaussians: Modeling dynamic urban scenes with Gaussian splatting,
[109] Zhou H, Shao J, Xu L, Bai D, Qiu W, Liu B, et al. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 21336–45.
[110] Johari MM, Carta C, Fleuret F. Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 17408–19.
[111] Li M, He J, Jiang G, Wang H. Ddn-slam: Real-time dense dynamic neural implicit slam with joint semantic encoding. 2024, arXiv preprint arXiv:2401.01545.
[112] Deng T, Wang Y, Xie H, Wang H, Wang J, Wang D, et al. Neslam: Neural implicit mapping and self-supervised feature tracking with depth completion and denoising. 2024, arXiv preprint arXiv:2403.20034.
[113] Yugay V, Li Y, Gevers T, Oswald MR. Gaussian-slam: Photo-realistic dense slam with gaussian splatting. 2023, arXiv preprint arXiv:2312.10070.
[114] Lu G, Zhang S, Wang Z, Liu C, Lu J, Tang Y. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In: European conference on computer vision. Springer; 2025, p. 349–66.
[115] Ze Y, Yan G, Wu Y-H, Macaluso A, Ge Y, Ye J, et al. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In: Conference on robot learning. PMLR; 2023, p. 284–301.
[116] Li B, Weinberger KQ, Belongie S, Koltun V, Ranftl R. Language-driven semantic segmentation. 2022, arXiv preprint arXiv:2201.03546.
[117] Shorinwa O, Tucker J, Smith A, Swann A, Chen T, Firoozi R, et al. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. In: 8th annual conference on robot learning. 2024.
[118] Shen W, Yang G, Yu A, Wong J, Kaelbling LP, Isola P. Distilled feature fields enable few-shot language-guided manipulation. 2023, arXiv preprint arXiv:2308.07931.
[119] Westover LA. Splatting: a parallel, feed-forward volume rendering algorithm. The University of North Carolina at Chapel Hill; 1991.
[120] Radl L, Steiner M, Parger M, Weinrauch A, Kerbl B, Steinberger M. Stopthepop: Sorted gaussian splatting for view-consistent real-time rendering. ACM Trans Graph 2024;43(4):1–17.
[121] Feng Q, Cao G, Chen H, Mu T-J, Martin RR, Hu S-M. A new split algorithm for 3D Gaussian splatting. 2024, arXiv preprint arXiv:2403.09143.
[122] Wei M, Wu Q, Zheng J, Rezatofighi H, Cai J. Normal-GS: 3D Gaussian splatting with normal-involved rendering. 2024, arXiv preprint arXiv:2410.20593.
[123] Chen H, Chen R, Qu Q, Wang Z, Liu T, Chen X, et al. Beyond Gaussians: Fast and high-fidelity 3D splatting with linear kernels. 2024, arXiv preprint arXiv:2411.12440.
[124] Zhang J, Zhan F, Xu M, Lu S, Xing E. Fregs: 3d gaussian splatting with progressive frequency regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 21424–33.
[125] Yan Z, Low WF, Chen Y, Lee GH. Multi-scale 3d gaussian splatting for anti-aliased rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 20923–31.
[126] Bi Z, Zeng Y, Zeng C, Pei F, Feng X, Zhou K, et al. Gsˆ 3: Efficient relighting with triple Gaussian splatting. 2024, arXiv preprint arXiv:2410.11419.
[127] Bayesian spatial kernel smoothing for scalable dense semantic mapping, author=gan, lu and zhang, ray and grizzle, jessy w and eustice, ryan m and ghaffari, maani. IEEE Robot Autom Lett 2020;5(2):790–7.
[128] Qiu R-Z, Sun Y, Marques JMC, Hauser K. Real-time semantic 3D reconstruction for high-touch surface recognition for robotic disinfection. In: 2022 IEEE/RSJ international conference on intelligent robots and systems. IEEE; 2022, p. 9919–25.
[129] Li Z, Zhao J, Zhou X, Wei S, Li P, Shuang F. RTSDM: A real-time semantic dense mapping system for UAVs. Machines 2022;10(4):285.
[130] Chang Y, Tian Y, How JP, Carlone L. Kimera-multi: a system for distributed multi-robot metric-semantic simultaneous localization and mapping (SLAM). In: 2021 IEEE international conference on robotics and automation. IEEE; 2021, p. 11210–8.
[131] Li M, Liu S, Zhou H. SGS-SLAM: Semantic Gaussian splatting for neural dense SLAM. 2024, arXiv preprint arXiv:2402.03246.
[132] Ji Y, Liu Y, Xie G, Ma B, Xie Z. NEDS-SLAM: A novel neural explicit dense semantic SLAM framework using 3D Gaussian splatting. 2024, arXiv preprint arXiv:2403.11679.
[133] Kim J, Chung D, Kim Y, Kim H. Deep learning-based 3D reconstruction of scaffolds using a robot dog. Autom Constr 2022;134:104092.
[134] Kopácsi L, Baffy B, Baranyi G, Skaf J, Sörös G, Szeier S, et al. Cross-viewpoint semantic mapping: Integrating human and robot perspectives for improved 3D semantic reconstruction. Sensors 2023;23(11):5126.
[135] Yue Y, Zhao C, Li R, Yang C, Zhang J, Wen M, et al. A hierarchical framework for collaborative probabilistic semantic mapping. In: 2020 IEEE international conference on robotics and automation. IEEE; 2020, p. 9659–65.
[136] Guo X, Hu J, Chen J, Deng F, Lam TL. Semantic histogram based graph matching for real-time multi-robot global localization in large scale environment. IEEE Robot Autom Lett 2021;6(4):8349–56.
[137] Deng Y, Wang M, Yang Y, Yue Y. HD-CCSOM: Hierarchical and dense collaborative continuous semantic occupancy mapping through label diffusion. In: 2022 IEEE/RSJ international conference on intelligent robots and systems. IEEE; 2022, p. 2417–22.
[138] Yang G, Wang X, Zhu D, Li Y. Real-time dense 3D semantic mapping using RGB-D camera. In: 2023 42nd Chinese control conference. IEEE; 2023, p. 4419–24.
[139] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. Int J Robot Res (IJRR) 2013.
[140] Hackel T, Savinov N, Ladicky L, Wegner JD, Schindler K, Pollefeys M. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. In: ISPRS annals of the photogrammetry, remote sensing and spatial information sciences, vol. IV-1-W1, 2017, p. 91–8.
[141] Liu X, Lei J, Prabhu A, Tao Y, Spasojevic I, Chaudhari P, et al. SlideSLAM: Sparse, lightweight, decentralized metric-semantic SLAM for multi-robot navigation. 2024, arXiv preprint arXiv:2406.17249.
[142] Zhu S, Qin R, Wang G, Liu J, Wang H. SemGauss-SLAM: Dense semantic Gaussian splatting SLAM. 2024, arXiv preprint arXiv:2403.07494.
[143] Romakin V. Functional-voxel modeling of navigation algorithm ORCA. 2020.
[144] Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proc. computer vision and pattern recognition (CVPR), IEEE. 2017.
[145] Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, et al. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9297–307.
[146] Song S, Lichtenberg SP, Xiao J. SUN RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 567–76.
[147] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE; 2012, p. 3354–61.
[148] Hepp B, Nießner M, Hilliges O. Plan3D: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. ACM Trans Graph 2018;38(1):1–17.
[149] Cheng M-L, Matsuoka M, Liu W, Yamazaki F. Near-real-time gradually expanding 3D land surface reconstruction in disaster areas by sequential drone imagery. Autom Constr 2022;135:104105.
[150] Sucar E, Wada K, Davison A. NodeSLAM: Neural object descriptors for multi-view shape reconstruction. In: 2020 international conference on 3D vision (3DV). IEEE; 2020, p. 949–58.
[151] Yang J, Gao Y, Li D, Waslander SL. Robi: A multi-view dataset for reflective objects in robotic bin-picking. In: 2021 IEEE/RSJ international conference on intelligent robots and systems. IEEE; 2021, p. 9788–95.
[152] Song S, Kim D, Choi S. View path planning via online multiview stereo for 3D modeling of large-scale structures. IEEE Trans Robot 2021;38(1):372–90.
[153] Wang H, Huang Y, Zhang G, Rong Y. A novel method for dense point cloud reconstruction and weld seam detection for tubesheet welding robot. Opt Laser Technol 2023;163:109346.
[154] Wang S, Leroy V, Cabon Y, Chidlovskii B, Revaud J. Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 20697–709.
[155] He X, Sun J, Wang Y, Peng S, Huang Q, Bao H, et al. Detector-free structure from motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 21594–603.
[156] Wang J, Karaev N, Rupprecht C, Novotny D. Vggsfm: Visual geometry grounded deep structure from motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 21686–97.
[157] Wen B, Tremblay J, Blukis V, Tyree S, Müller T, Evans A, et al. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 606–17.
[158] Peng J, Xu W, Liang B, Wu A-G. Virtual stereovision pose measurement of noncooperative space targets for a dual-arm space robot. IEEE Trans Instrum Meas 2019;69(1):76–88.
[159] Cai Y, Ou Y, Qin T. Improving SLAM techniques with integrated multi-sensor fusion for 3D reconstruction. Sensors 2024;24(7):2033.
[160] Islam QU, Ibrahim H, Chin PK, Lim K, Abdullah MZ. MVS-SLAM: Enhanced multiview geometry for improved semantic RGBD SLAM in dynamic environment. J Field Robot 2024;41(1):109–30.
[161] Yin M, Wu S, Han K. IBD-SLAM: Learning image-based depth fusion for generalizable SLAM. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 10563–73.

[162] Niemeyer M, Mescheder L, Oechsle M, Geiger A. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 3504–15.
[163] Furrer F, Burri M, Achtelik M, Siegwart R. Rotors—a modular gazebo mav simulator framework. In: Robot operating system (ROS) the complete reference (volume 1). Springer; 2016, p. 595–625.
[164] Jensen R, Dahl A, Vogiatzis G, Tola E, Aanæ s H. Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, p. 406–13.
[165] Schops T, Schonberger JL, Galliani S, Sattler T, Schindler K, Pollefeys M, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 3260–9.
[166] Magistri F, Marks E, Nagulavancha S, Vizzo I, Läebe T, Behley J, et al. Contrastive 3D shape completion and reconstruction for agricultural robots using RGB-D frames. IEEE Robot Autom Lett 2022;7(4):10120–7.
[167] Xu Z, He Z, Wu J, Song S. Learning 3D dynamic scene representations for robot manipulation. In: Conference on robot learning (CoRL). 2020.
[168] Lundell J, Verdoja F, Kyrki V. DDGC: Generative deep dexterous grasping in clutter. IEEE Robot Autom Lett 2021;6(4):6899–906.
[169] Gao W, Tedrake R. KPAM-SC: Generalizable manipulation planning using key-point affordance and shape completion. In: 2021 IEEE international conference on robotics and automation. IEEE; 2021, p. 6527–33.
[170] Ge Y, Xiong Y, From PJ. Symmetry-based 3D shape completion for fruit localisation for harvesting robots. Biosyst Eng 2020;197:188–202.
[171] He J-Z, Wang D-J, Fang H, Fu Q-L, Zhou D-M. Inhibited transport of graphene oxide nanoparticles in granular quartz sand coated with Bacillus subtilis and Pseudomonas putida biofilms. Chemosphere 2017;169:1–8.
[172] Gualtieri M, Platt R. Robotic pick-and-place with uncertain object instance segmentation and shape completion. IEEE Robot Autom Lett 2021;6(2):1753–60.
[173] Qin Y, Chen R, Zhu H, Song M, Xu J, Su H. S4G: Amodal single-view single-shot se(3) grasp detection in cluttered scenes. In: Conference on robot learning. PMLR; 2020, p. 53–65.
[174] Mohammadi SS, Duarte NF, Dimou D, Wang Y, Taiana M, Morerio P, et al. 3DSgrasp: 3D shape-completion for robotic grasp. In: 2023 IEEE international conference on robotics and automation. IEEE; 2023, p. 3815–22.
[175] Calli B, Walsman A, Singh A, Srinivasa S, Abbeel P, Dollar AM. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. IEEE Robot Autom Mag 2015;22(3):36–52.
[176] Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, et al. ShapeNet: An information-rich 3D model repository. 2015, arXiv preprint arXiv:1512.03012.
[177] Agnew W, Xie C, Walsman A, Murad O, Wang C, Domingos P, et al. Amodal 3D reconstruction for robotic manipulation via stability and connectivity. In: proceedings of the 2024 conference on robot learning. PMLR; 2024, URL https://proceedings.mlr.press/v155/agnew21a.html.
[178] Zhan G, Zheng C, Xie W, Zisserman A. Amodal ground truth and completion in the wild. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 28003–13.
[179] Calli B, Singh A, Walsman A, Srinivasa S, Abbeel P, Dollar AM. The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics. IEEE; 2015, p. 510–7.
[180] Bian W, Wang Z, Li K, Bian J-W, Prisacariu VA. Nope-nerf: Optimising neural radiance field with no pose prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 4160–9.
[181] Liu J, Nie Q, Liu Y, Wang C. Nerf-loc: Visual localization with conditional neural radiance field. In: 2023 IEEE international conference on robotics and automation. IEEE; 2023, p. 9385–92.
[182] Kerr J, Fu L, Huang H, Avigal Y, Tancik M, Ichnowski J, et al. Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In: 6th annual conference on robot learning. 2022.
[183] Ichnowski J, Avigal Y, Kerr J, Goldberg K. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In: Conference on robot learning. PMLR; 2022, p. 526–36.
[184] Zhou A, Kim MJ, Wang L, Florence P, Finn C. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 17907–17.
[185] Dai Q, Zhu Y, Geng Y, Ruan C, Zhang J, Wang H. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In: 2023 IEEE international conference on robotics and automation. IEEE; 2023, p. 1757–63.
[186] Ze Y, Yan G, Wu Y-H, Macaluso A, Ge Y, Ye J, et al. Multi-task real robot learning with generalizable neural feature fields. In: 7th annual conference on robot learning. 2023.
[187] Adamkiewicz M, Chen T, Caccavale A, Gardner R, Culbertson P, Bohg J, et al. Vision-only robot navigation in a neural radiance world. IEEE Robot Autom Lett 2022;7(2):4606–13.
[188] Byravan A, Humplik J, Hasenclever L, Brussee A, Nori F, Haarnoja T, et al. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields. In: 2023 IEEE international conference on robotics and automation. IEEE; 2023, p. 9362–9.
[189] Lee S, Chen L, Wang J, Liniger A, Kumar S, Yu F. Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. IEEE Robot Autom Lett 2022;7(4):12070–7.
[190] Chen X, Li Q, Wang T, Xue T, Pang J. GenNBV: Generalizable next-best-view policy for active 3D reconstruction. 2024, arXiv preprint arXiv:2402.16174.
[191] Wang Y, Yan Y, Shi D, Zhu W, Xia J, Jeff T, et al. NeRF-IBVS: visual servo based on NeRF for visual localization and navigation. Adv Neural Inf Process Syst 2024;36.
[192] Jiang Z, Zhu Y, Svetlik M, Fang K, Zhu Y. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. 2021, arXiv preprint arXiv:2104.01542.
[193] Wu T, Zhang J, Fu X, Wang Y, Ren J, Pan L, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 803–14.
[194] Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, VanderBilt E, et al. Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 13142–53.
[195] Durvasula S, Kiguru R, Mathur S, Xu J, Lin J, Vijaykumar N. VoxelCache: Accelerating online mapping in robotics and 3D reconstruction tasks. In: Proceedings of the international conference on parallel architectures and compilation techniques. 2022, p. 239–51.
[196] Yamazaki K, Hanyu T, Vo K, Pham T, Tran M, Doretto G, et al. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In: 2024 IEEE international conference on robotics and automation. IEEE; 2024, p. 9411–7.
[197] Li XS, Nguyen T, Cohn AG, Dogar M, Cohen N. Real-time robot topological localization and mapping with limited visual sampling in simulated buried pipe networks. Front Robot AI 2023;10:1202568.
[198] Huang C, Mees O, Zeng A, Burgard W. Visual language maps for robot navigation. In: Proceedings of the IEEE international conference on robotics and automation. London, UK; 2023.
[199] Chen T, Culbertson P, Schwager M. Catnips: Collision avoidance through neural implicit probabilistic scenes. IEEE Trans Robot 2024.
[200] Pumarola A, Corona E, Pons-Moll G, Moreno-Noguer F. D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 10318–27.
[201] Šlapak E, Pardo E, Dopiriak M, Maksymyuk T, Gazda J. Neural radiance fields in the industrial and robotics domain: applications, research opportunities and use cases. Robot Comput-Integr Manuf 2024;90:102810.
[202] Bhatnagar BL, Xie X, Petrov IA, Sminchisescu C, Theobalt C, Pons-Moll G. BEHAVE: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 15935–46.
[203] Kasper A, Xue Z, Dillmann R. The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. Int J Robot Res 2012;31(8):927–34.
[204] Hampali S, Rad M, Oberweger M, Lepetit V. HOnnotate: A method for 3D annotation of hand and object poses. In: CVPR. 2020.
[205] Wen B, Mitash C, Ren B, Bekris KE. Se(3)-TrackNet: Data-driven 6D pose tracking by calibrating image residuals in synthetic domains. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE; 2020, http://dx.doi.org/10.1109/iros45743.2020.9341314.
[206] Morrison D, Corke P, Leitner J. EGAD! An evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. IEEE Robot Autom Lett 2020;5(3):4368–75.
[207] Smitt C, Halstead M, Zaenker T, Bennewitz M, McCool C. Pathobot: A robot for glasshouse crop phenotyping and intervention. In: 2021 IEEE international conference on robotics and automation. IEEE; 2021, p. 2324–30.
[208] Mahler J, Liang J, Niyaz S, Laskey M, Doan R, Liu X, et al. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. 2017, arXiv preprint arXiv:1703.09312.
[209] Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D SLAM systems. In: Proc. of the international conference on intelligent robot systems. 2012.
[210] Petrovai A, Nedevschi S. Semantic cameras for 360-degree environment perception in automated urban driving. IEEE Trans Intell Transp Syst 2022;23(10):17271–83.
[211] Barron JT, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R, Srinivasan PP. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 5855–64.
[212] Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, et al. The EuRoC micro aerial vehicle datasets. Int J Robot Res 2016. http://dx.doi.org/10.1177/0278364915620033, arXiv:http://ijr.sagepub.com/content/early/2016/01/21/0278364915620033.full.pdf+html. URL http://ijr.sagepub.com/content/early/2016/01/21/0278364915620033.abstract.

[213] Straub J, Whelan T, Ma L, Chen Y, Wijmans E, Green S, et al. The replica dataset: A digital replica of indoor spaces. 2019, arXiv preprint arXiv:1906.05797.

[214] Carlevaris-Bianco N, Ushani AK, Eustice RM. University of Michigan North Campus long-term vision and lidar dataset. Int J Robot Res 2015;35(9):1023–35.

[215] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The CityScapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 3213–23.

[216] Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: The IEEE conference on computer vision and pattern recognition. 2016.

[217] Shah S, Dey D, Lovett C, Kapoor A. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In: Field and service robotics. 2017, arXiv:arXiv:1705.05065. URL https://arxiv.org/abs/1705.05065.

[218] Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, et al. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In: Proc. of the IEEE/CVF international conf. on computer vision. 2019.

[219] Jiang L, Zhao H, Shi S, Liu S, Fu C-W, Jia J. Pointgroup: Dual-set point grouping for 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 4867–76.

[220] Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Scene parsing through ADE20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 633–41.