Toward Semantic Scene Understanding: Benchmarking for Mobile Robot Navigation Indoors

Isaac Asante, Lau Bee Theng(✉), and Mark Tee Kit Tsun

Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia
blau@swinburne.edu.my

Abstract. Mobile robot navigation is a constantly evolving field that is adopting new paradigms along the way, and recent methods, such as Transformer-based models, have helped facilitate advancements in perception and decision-making tasks in this decade alone. This paper explores modern scene understanding techniques, including Contrastive Language-Image Pretraining (CLIP) and its role in improving semantic scene comprehension for various indoor environments. Existing benchmarking methods for evaluating autonomous mobile robot navigation performance are limited in accommodating the dynamic nature of real-world scenarios. Therefore, a set of metrics is proposed for robust evaluation, highlighting the need for standardized frameworks that meet modern expectations. Furthermore, a multimodal robot navigation model is introduced; it consists of visual and laser data combined with semantic embeddings to augment navigation performance. The proposed model and metrics aim to contribute to better benchmarking standards for indoor robot navigation systems.

Keywords: Autonomous Mobile Robots · Indoor Navigation · Scene Understanding · Benchmarking Standards · Contrastive Language-Image Pretraining · Metrics

1 Introduction

Transformer-based models have become more prevalent in recent years and have inevitably received widespread adoption in mobile robotics – whether in the industry or academic research – where artificial intelligence (AI) is heavily used to solve problems at different scales. They were first introduced in the last decade to augment natural language processing (NLP) systems [1], and today, their architecture is leveraged to aid robots in perceiving and navigating environments thanks to their ability to handle long-term dependencies and suitability for complex perception tasks, including advanced object detection. Notable advancements include 3DETR [2], which predicts 3D bounding boxes from point cloud data. Voxel Set Transformer (VoxSet) [3] and the earlier Voxel Transformer (VoTr) [4] can process sparse point clouds, whereas the Vision Transformer (ViT) contributes to monocular navigation via image segmentation [5].

D.-J. Deng and J.-C. Chen (Eds.): SGIoT 2024, LNICST 633, pp. 90–106, 2026.
https://doi.org/10.1007/978-3-031-93825-2_9

Scene understanding involves processing visual and spatial information, which remains an active research area [6]. Room type classification [7] is a common task in that field, as it tests the ability to comprehend spatial relationships across varying layouts and dimensions [8]. However, in autonomous mobile robot navigation systems developed for real-world dynamic indoor environments, scene understanding extends to semantic comprehension, considering several factors beyond geometric shapes, proximities, and immediate surroundings to derive their context and functions. Contrastive learning holds much potential [9, 10], although other methods exist to achieve the desired commonsense reasoning, and many solutions still revolve around Transformers [11].

In subsequent sections, this paper explores modern scene understanding techniques. It focuses on Contrastive Language-Image Pretraining (CLIP) [9], its benefits, and how it can be a foundational block in other scene perception models. This paper also addresses the lack of standardization in current benchmarking datasets and frameworks, which tend to be limited when evaluating end-to-end robot navigation performance according to modern expectations. The discussion affirms that modern research is shifting towards semantic scene understanding in dynamic indoor environments, and traditional benchmarks are becoming inadequate as they were either designed for static settings and simpler navigation tasks or they failed to share similar structures across the field. In response, a list of metrics is proposed for appropriate navigation system evaluations. Moreover, a multimodal robot navigation model is introduced, which uses visual and laser data and semantic embeddings for advanced navigation performance. It can be tested in the future according to the proposed metrics.

2 Scene Understanding via Contrastive Learning

Task-agnostic zero-shot classification may be considered one of the prerequisites to extract information from an environment observed through a camera more efficiently. Yet this area of research has been understudied in recent years, as stressed by OpenAI in their 2021 study [9] when they proposed CLIP, a model designed to connect visual information with natural language descriptions with the zero-shot approach through contrastive learning. Instead of a traditional supervised learning technique involving a convolutional neural network (CNN) as the backbone for feature extraction and mapping features to a pre-determined set of categories, CLIP employs two encoders: one for images, which may be a Vision Transformer and one for text captions. The rationale is based on the observation that the restricted form of traditional supervision limits model generality and usability with unseen object categories.

CLIP learns a joint embedding space encoding representations of visual and text modalities, thus becoming able to perform various vision-language functions with direct carryovers to mobile robot navigation. The underlying principle is that mapping a text prompt into the embedding space to identify an image close to the feature space is an effective method of obtaining a high degree of semantic similarity. Thus, CLIP is configured during training to maximize the cosine similarity of image and text embeddings of matching text-image pairs in the training batch while minimizing the similarity of pairs with semantically misaligned components. However, this contrastive learning alone is insufficient to guarantee viable application in new datasets – or in a robotics

context, unseen environments – without task-specific tuning. The study takes note of the unsuitability of standard crowd-labelled datasets like Microsoft COCO [12] or Visual Genome [13] with roughly 100,000 photos, which pales in comparison to CLIP's 400 million image-text pairs covering 500,000 queries, each of which addresses a distinct visual concept. While this may not be entirely relevant in a mobile robot navigation prototype for dynamic indoor environments, it emphasizes the need to transition to more modern training datasets and input types to solve modern problems better. Linking back to robotics, autonomous mobile robot navigation in uncontrolled, real-world indoor settings featuring small ground obstacles, for instance, is likely to present a combination of unexpected elements, forming scenarios that are difficult for the robot to understand and negatively impact its path planning or behaviour [14]. Therefore, achieving proper scene understanding via extracting semantic information in the real world depends on the underlying visual information processing model's zero-shot capabilities. The model used must be built through a training dataset that is large and diverse enough to guarantee a high rate of sensible, accurate inferences even when faced with novel scenarios. Upgrading the size and nature of the standard training datasets is an adequate step in that direction.

This work draws attention to the relationship of perceived elements in an environment, advocating for a paradigm shift for less naïve, more intuitive computer vision models. It goes beyond specific tasks such as image caption generation, which the CLIP development team measured as more computationally demanding and difficult to scale. They note that a 63 million parameter Transformer language model consumed twice as much computational resources as a ResNet-50 image encoder, revealing scalability issues when creating a model similar to VirTex [15]. Nevertheless, more recent research [16] has offered a different perspective, and evaluations show that captioning models may outperform contrastive models in few-shot classification and fine-grained tasks. Yet, CLIP remains a superior option for zero-shot classification.

With the quest for Artificial General Intelligence (AGI) still ongoing, the properties of natural intelligence reflecting human cognitive processes and behaviour – blending analytic, creative and practical intelligence [17] – continuously emerge into other technologies as researchers rethink fundamental concepts of their fields for more robust and adaptive problem-solving methods. In 2023, a collaborative effort between members of various institutions, including Google Research and Waymo LLC, proposed OpenScene, a novel method for answering advanced queries that may aid a robot to interact intelligently with and understand a scene [18]. The novelty is that the model addresses semantics, materials, room types, affordances, functions and physical properties of every 3D point captured without requiring explicit 3D annotations for those concepts. It projects 2D features to 3D points using camera poses and depth information, training the MinkowskiNet18A3D network [19] using cosine similarity loss to align 3D embeddings with CLIP's feature space. Ultimately, this enables open-vocabulary scene understanding by representing 3D points in terms of broader visual concepts. Incidentally, the semantic image segmentation required here may be performed via OpenSeg [20] or LSeg [21]; both come with open-vocabulary capabilities and improved generalization.

It demonstrates scene understanding beyond object labels from predefined categories. It suggests that this applicability may enable a navigation robot to recognize and prioritize

obstacles in a dynamic environment by nature, whether mobile or static and potentially by danger levels, even though the study does not provide enough material to support or refute the latter hypothesis. Yet, that cumulative knowledge may aid in focusing resources on visible areas where path planning may be safer by transferring semantic data and inferences across sensory input readings to make better navigation decisions.

It remains to be seen whether OpenScene's computational efficiency and real-time performance suit active autonomous robot navigation. Nonetheless, such an omitted factor reflects current research gaps regarding mobile robot scene understanding based on semantic information.

The dissimilarity between modern techniques for robot scene perceptions and human perception capabilities is commonly mentioned when key problems are explored [22], such as limited sensory modalities, over-reliance on pre-trained models, lack of adaptability, and lack of commonsense reasoning in unstructured, real-world environments [23, 24]. Other limitations have emerged in recent studies, such as the lack of standardized benchmarks and evaluation metrics for semantic mapping systems regarding sensor types and mounted positions. [25]. There is also the inability to reflect the complexities of real-world scenarios indoors where robots are compelled to navigate in the presence of moving objects and people, a key attribute of a dynamic environment [26]. Additionally, there are challenges in complex scenes, such as dynamic objects, changing lighting, and cluttered environments, where multi-sensor fusion involving 3D cameras, IMU sensors, and even laser distances may be necessary [22].

3 Lack of Benchmarking Standards in Modern Literature

Autonomous mobile robot navigation involves several subdomains requiring distinct methodologies and evaluation metrics. It includes environment perception using single or multiple sensors, object detection, mapping, localization, path planning, and locomotion [27]. For socially aware robots, cognitive functions are important to assess, counting semantic scene understanding and context reasoning [14, 28], which help with decision-making and interactions. The choice of navigation methods implemented and tests performed are easily influenced by the type of environment, whether indoor, outdoor, static or dynamic [29]. Dynamic indoor environments, for instance, are characterized by moving obstacles with varying properties in modern literature [30]. Relevant examples include restaurants or supermarkets with pedestrians, shopping carts, or other dynamic objects [26] that compel an autonomous wheeled agent to have robust spatial and contextual awareness to avoid collisions during navigation. In such cases, specifying the type of mobile objects, their moving speed, and starting positions is essential before evaluations as they directly affect the robot's real-time path planning and decision-making processes. Thus, standardized evaluation frameworks with predefined metrics and test cases are vital for benchmarking by researchers.

Over 910 companies worldwide are confirmed to use the Robot Operating System (ROS) for their robotics platforms, and ROS 2 packages were downloaded over 300 million times in 2023 alone [31], which includes uses in academic studies, making the operating system an industry standard. Yet, its prominent navigation framework, *Nav2* [32], remains underused in literature, and many of its algorithms have yet to be

benchmarked against other state-of-the-art navigation solutions outside of the ROS 2 ecosystem [33].

Today, benchmarking datasets and frameworks continue to be developed to address limitations noted in existing systems. The M2DGR dataset [34] was released in 2021 by Shanghai Jiao Tong University researchers to simplify the development and evaluation of Simultaneous Localization and Mapping (SLAM) algorithms for ground robots. It includes multi-sensor data captured using a fisheye, standard RGB, event and infrared cameras, 32-beam LiDAR for 3D point clouds, inertial sensors, and a consumer-grade GNSS receiver. In addition to outdoor scenarios and transition sequences, its indoor environment scenarios consist of corridors, halls, lifts, and rooms and scenarios in complete darkness. Dynamic motion sequences are added to evaluate cases requiring rapid turns due to zigzag routes or situations prompting speed changes. One of the primary metrics used in M2DGR is the Absolute Trajectory Error (ATE) [35], calculated by the EVO tool [36] designed to evaluate and compare trajectory outputs of SLAM algorithms. Upon testing existing state-of-the-art SLAM algorithms on M2GDR, notable failures were recorded. The pinhole and fisheye versions of the visual SLAM method ORB-SLAM3 [37] failed their tracking and mapping in the dark environment from the *Roomdark06* sequence. This testing exposes limitations in certain cutting-edge SLAM techniques in low-illumination conditions, where adding LiDAR can provide the necessary depth information to compensate for the indistinguishable visual features.

Conversely, LiDAR-based systems A-LOAM [38], LeGO-LOAM [39], LINS [40], and LIO-SAM [41] all struggled with the *Lift04* sequence, and none successfully reconstructed a complete map. Inaccuracies in elevator transitions stem from the fact that the interior may appear featureless or static for LiDAR scans. In contrast, IMUs record the elevator's vertical motion, thus causing discrepancies when reconciling the data between the different sensor modalities.

From OpenAI's assertion that older datasets such as MS-COCO are insufficient for developing modern zero-shot classification models [9] to M2DGR's efforts to address significant gaps in SLAM algorithm evaluations for ground robots [34], a pattern emerges: modern literature highlights the shortcomings of existing standard datasets for addressing new challenges and opportunities. These limitations span the entire process in mobile robotics and related fields, from model training to benchmarking; hence, standardized metrics and test cases are needed to improve representativeness, address edge cases and encompass greater diversity and scale. Table 1 below summarises the limitations of existing standard RGB-D 2D and 3D datasets for benchmarking mobile robot navigation in dynamic indoor environments, per the work in [26].

The THUD dataset [26] presents Mean Average Precision (mAP), Mean Intersection over Union (MIoU), Translation Error and Rotation Error as benchmarking metrics to measure robot navigation performance and scene understanding in dynamic indoor spaces. The mAP evaluates 3D object detection algorithms, MIoU assesses semantic segmentation accuracy, whereas the translation and rotation errors are recorded to gauge robot re-localization. The impact of the density of moving objects on robot re-localization is quantified via Dynamic Complexity, although the latter is not considered a direct performance metric.

Table 1. Limitations of existing datasets for benchmarking of mobile robot navigation.

Dataset	Description	Limitations
B3DO [42]	Provides 2D bounding box annotations on RGB-D images, limited frames (849) and object classes (50+)	No 3D annotations; lacks dynamic objects
NYU-Depth v2 [43]	It includes 2D semantic segmentation from RGB-D videos, a small number of annotated frames (1,449), and object classes (894)	No 3D annotations; lacks dynamic objects
SUN3D [44]	Comprises 415 RGB-D video sequences across 254 scenes and lacks dense annotations and object class information	Sparse annotations (10–15 per frame); no 3D annotations or dynamic objects
Stanford 2D-3D-S [45]	It offers large-scale virtual scenes with 2D texture, geometry, and semantic info; it relies on iGibson simulation	Relies on iGibson simulation and lacks dynamic objects
SceneNet RGB-D [46]	Provides 5 million photorealistic images of synthetic indoor trajectories and various 2D annotations	Lacks 3D annotations and dynamic objects
SUN RGB-D [47]	Contains 10,335 RGB-D images with dense 2D/3D annotations, including 2D polygons, 3D bounding boxes, and room layouts	Lacks dynamic objects and pedestrians
ScanNet [48]	Comprises 1,513 video sequences with 3D camera poses, surface reconstruction, semantic segmentation, and CAD models	Lacks dynamic objects and pedestrians
SUN-CG [49]	Provides 45,000 virtual scene layouts and 500,000 rendered images with single-view RGB, depth maps, and segmentation maps	No dynamic object annotations
Matterport 3D [50]	Contains 194,400 RGB-D images for generating panoramas with surface reconstruction, camera positions, and 2D/3D annotations	No dynamic object annotations

(*continued*)

Table 1. (*continued*)

Dataset	Description	Limitations
InteriorNet [51]	Rendered in virtual home scenes containing 15,000 sequences	No dynamic object annotations
ARKitScenes [52]	Improves ground truth geometry resolution from laser scans	No dynamic object annotations

The 3RScan dataset [53] offers value in object instance re-localization, an important aspect of end-to-end autonomous robot navigation. 3RScan was built from real-world scans of indoor environments using an RGB-D camera and is a viable option to support research and benchmarking in 3D scene understanding where objects move as part of a dynamic environment. Nonetheless, benchmarking navigation performance requires appropriate quantitative metrics. Suitable metrics include path length to determine efficiency in reaching a set goal, navigation time, number of collisions to assess obstacle avoidance capabilities, and success rate in completing assigned navigation tasks.

In 2024, Rondoni et al. set out to develop benchmarking metrics for autonomous medical robot navigation in the presence of both static and dynamic obstacles [54]. The motivation was that despite the importance of consistent, reliable, safe navigation in mobile robotics, there was no standardized approach for evaluating robot navigation in hospitals. For example, the safety requirements and guidelines for medical robots in the ISO 13482:2014 standard [55, p. 13482] do not include explicit benchmarks of structured tests on defined platforms for navigation performance. This lack of specificity leaves room for researchers and developers to define their benchmarking metrics, leading to fragmented and inconsistent evaluation methods. Therefore, it becomes difficult to benchmark and compare modern autonomous robot navigation systems.

The researchers [54] proposed several standardized tests in a simulated hospital environment to address this issue. Four batches of tests are presented in the study.

- Batch 1 consists of an obstacle-free path, portraying the simplest navigating condition.
- Batch 2 involves navigation around one or two static square obstacles of varying sizes (0.03 m, 0.15 m, 0.30 m, and 0.60 m) and placements – in parallel or perpendicular – with double obstacle configurations spaced at varying distances.
- Batch 3 includes interaction with a teleoperated robot programmed to move at predefined linear and angular speeds to simulate passing, crossing, and overtaking conditions.
- Batch 4 combines static and moving obstacles.

The selected test speeds for the robots – 0.2 m/s, 0.6 m/s, and 1.0 m/s – were meant to mirror typical indoor human walking speeds. Seven key metrics were recorded, as shown in Fig. 1: completion time, path length, distance error, orientation error, success rate, minimum distance from obstacles, and time at cruise speed.

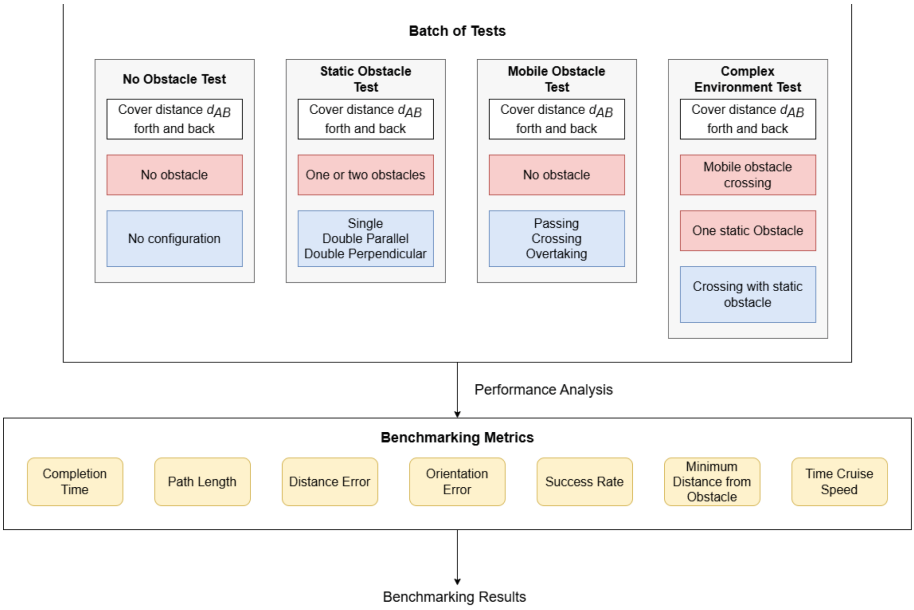

Fig. 1. Benchmarking protocol used in [54].

4 Proposed Benchmarking Metrics

Consolidating the information from explored studies in this paper, in addition to recent ROS-based benchmarking frameworks, such as Arena-Rosnav 2.0 [56] and HuNavSim [57], this work emphasizes that new benchmarking datasets must be developed to meet the necessary standards for human-aware autonomous mobile robot navigation in dynamic indoor environments with support for various sensor modalities. This paper proposes a list of 20 quantitative and qualitative metrics to evaluate navigation performance, semantic scene understanding, and social compliance. Table 2 below presents the proposed benchmarking metrics to serve as the standard for evaluating human-aware autonomous mobile robot navigation systems indoors.

Table 2. Proposed benchmarking metrics.

Category	Metric	Description
Navigation Performance	CompletionTime (CT)	Time taken to reach the goal (seconds)
	Path Length (PL)	Total distance traversed (meters)

(*continued*)

Table 2. (*continued*)

Category	Metric	Description
	Success Rate (SR)	Percentage of successful goal reaches (%)
	Deviation Error (DE)	Difference between the robot's position and the planned path (meters)
	Orientation Error (OE)	Difference between the robot's orientation and the planned path (degrees)
	Minimum Distance from Obstacle (MDO)	Closest proximity to obstacles during navigation (meters)
	Time Cruise Speed (TCS)	Percentage of time the robot maintains its desired speed (%)
	Collision Rate	Number of collisions with obstacles (count)
	Clearing Distance	Distance maintained from obstacles (meters)
	Movement Jerk	Smoothness of the robot's accelerations (meters/second^3)
	Trajectory Angle Over Length	Smoothness of the robot's path (degrees/meter)
	Trajectory Roughness	Variations in the robot's path (no unit)
Semantic Scene Understanding	Mean Average Precision (mAP)	Accuracy of 3D object detection based on bounding box predictions and class labels (%)
	Intersection over Union (IoU)	Overlap between predicted and ground truth bounding boxes for object detection (%)
	Semantic Mapping Accuracy	Accuracy and completeness of the semantic map, including the method of label assignment and its navigational impact
	Zero-Shot Object Recognition	The robot's ability to recognize and locate objects not present in its training set

(*continued*)

Table 2. (*continued*)

Category	Metric	Description
	Scene Classification Accuracy	Accuracy of the robot's classification of different scene types (e.g., 'office', 'hospital room') (%)
	Spatial Relationship Understanding	The robot's ability to recognize spatial relationships between objects (e.g., 'next to', 'behind') (no unit)
	Average Distance to Closest Person	Average distance maintained from the closest person (meters)
	Minimum Distance to People	Minimum distance maintained from people in the environment (meters)

5 Proposed Mobile Robot Navigation Model

Finally, to address the issues hitherto discussed, this work introduces the architecture of a model designed to improve mobile robot navigation in indoor environments with various obstacles, aligned with the benchmarking metrics listed in the previous section. The primary objective is to enhance scene perception through advanced semantic understanding for safer autonomous navigation. The proposed model addresses challenges explored in modern research, as seen in recent literature, which is increasingly focused on intuitive zero-shot capable systems. The purpose is to handle indoor settings with small static ground obstacles or unfamiliar objects, overcoming the limitations of older visual sensory processing models trained on standard computer vision datasets such as MS-COCO with predefined classes. Figure 2 and Fig. 3 demonstrate how the model adopts a multimodal approach to leverage the complementary strengths of different sensors – in this case, a 2D laser scanner and an RGB-D camera. Together, this allows for the collection of range and intensity data, which is useful for mapping and obstacle avoidance, as well as depth information plus RGB for 3D perception. The combined sensory input data is meant to be aligned and normalized to ensure they are in the same coordinate frame and scale at every timestamp and fused into a unified representation holding vital scene details.

Some components from the OpenScene pipeline [18] are reflected in this model's *Semantic Segmentation and Enrichment* module. OpenSeg generates pixel-wise object boundaries from the camera's RGB data. CLIP processes the segmentations to extract feature embeddings by mapping them into a semantic space that enriches the model's understanding of detected objects and contextual relationships within the scene, thus improving the robot's interpretation of the perceived environment. Integrating the segmentation masks and CLIP's embeddings into the fused sensory data – including scan angles, ranges, intensities, camera depth, RGB frames and derived blind spots – improves the model's semantic understanding and spatial awareness of the environment. This improvement is due to the additional object classification details that enhance

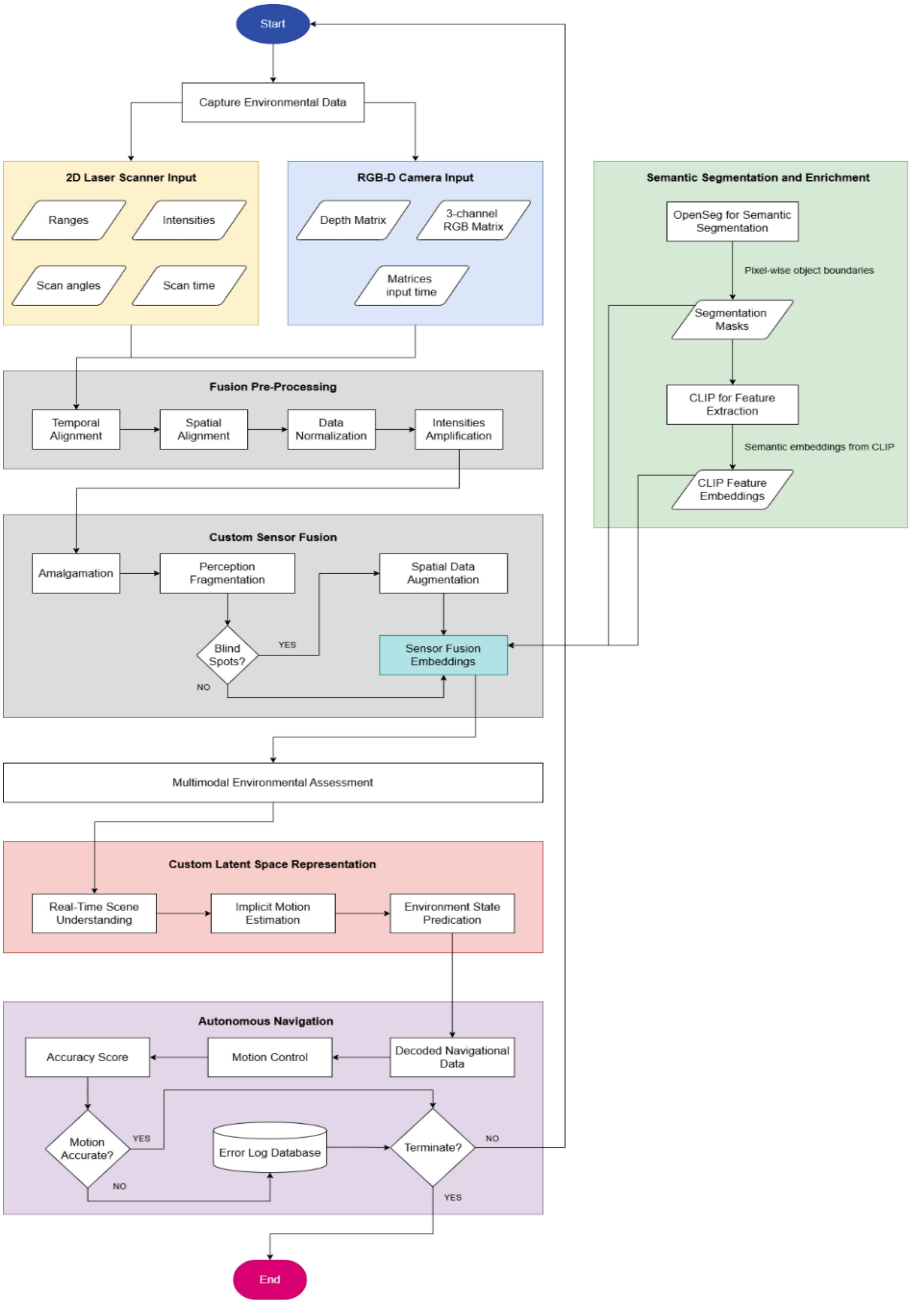

Fig. 2. Overview of the proposed model.

```
# Inputs:
# laser_scan      - Laser scan data (ranges)
# intensities     - Laser intensity values (already between 0 and 1)
# depth_matrix    - Depth data from RGB-D camera
# rgb_matrix      - 3-channel RGB data from the camera
# angle_min       - Minimum angle of the laser scan
# angle_max       - Maximum angle of the laser scan
# angle_increment- Angular increment between laser scan points
# matrices_time   - Input timestamps from sensors
# camera_fov      - Field of view of the RGB camera
# depth_intrinsics - Intrinsic parameters of the depth camera
# rgb_intrinsics    - Intrinsic parameters of the RGB camera
# extrinsics        - Extrinsic parameters between the depth and RGB cameras

# Align sensor data based on timestamps
temporal_alignment = align_temporal(laser_scan, depth_matrix, rgb_matrix,
matrices_time)

# 2. Spatial Calibration between depth and RGB
calibrated_depth_matrix =
calibrate_depth_to_rgb(depth_matrix, depth_intrinsics, rgb_intrinsics, ex-
trinsics)

# 3. Spatial Alignment of Laser Data to Camera Frame
spatially_aligned_laser = project_laser_to_camera(laser_scan, angle_min,
angle_max, angle_increment, camera_fov)
spatially_aligned_rgb = rgb_matrix  # RGB data is in its frame
spatially_aligned_depth = calibrated_depth_matrix  # Calibrated depth data
to match RGB FOV

# 4. Data Normalization
laser_normalized = normalize_laser(spatially_aligned_laser)
depth_normalized = normalize_depth(spatially_aligned_depth)

# 5. Amplify the intensity values
amplified_intensities = amplify_intensities(intensities, threshold=0.5)

# 6. Custom Sensor Fusion
amalgamated_data = combine_data(laser_normalized, depth_normalized, spatial-
ly_aligned_rgb, amplified_intensities)

# 7. Perception Fragmentation
left_blind_spot = calculate_blind_spot(angle_min, angle_max, an-
gle_increment, camera_fov, "left")
right_blind_spot = calculate_blind_spot(angle_min, angle_max, an-
gle_increment, camera_fov, "right")
left_blind_spot_data = extract_blind_spot_data(laser_scan, left_blind_spot)
right_blind_spot_data = extract_blind_spot_data(laser_scan,
right_blind_spot)

# 8. Spatial Data Augmentation
augmented_data = augment_data(amalgamated_data, left_blind_spot_data,
right_blind_spot_data)

# 9. Generate Sensor Fusion Embeddings
sensor_fusion_embeddings = fuse_sensor_data(augmented_data)

# Release fused embeddings for downstream modules

return sensor_fusion_embeddings
```

Fig. 3. Proposed algorithm for the sensor fusion preprocessing and custom sensor fusion.

spatial reasoning. It leads to the robot's proficiency in classifying objects of different kinds, delimitating shapes, estimating proximity, identifying free space around them, and improving overall contextual awareness. This operation enhances the mobile robot's ability to identify favourable navigation paths, detect potentially hazardous objects of any size and recognize small ground obstacles that could impact the robot's hardware components, especially in structured indoor settings.

This amalgamation creates a latent space representation of the environment that encodes spatial layouts, object identities, and proximity relationships due to transforming spatial and semantic embeddings at various stages of the process flow. The latest space representation provides essential information for collision avoidance, motion estimation, and path planning. It can also be decoded via custom neural network-based algorithms to generate the robot's angular and linear velocity commands. These motion commands are continuously scored through a feedback mechanism, such as calculating the collision risks or the difference in angles and distance to an assigned goal, to refine the robot's autonomous navigation accuracy. Each of the model's single components may be evaluated separately in inference mode or tested simultaneously as one unit, using the relevant benchmarks from Table 2.

6 Conclusion and Future Work

Several algorithms for semantic scene understanding in connection to mobile robot navigation were studied in this paper. A part of the literature review focused on contrastive learning techniques like the Transformer-powered CLIP model, which has several benefits for use in diverse indoor scenarios. The study found that capturing the nuances of a dynamic real-world indoor setting during evaluations is challenging due to the limitations in the present benchmarking frameworks, not to mention the absence of standardized test cases and metrics throughout the literature.

This research developed a set of 20 evaluation metrics tailored for human-aware autonomous robot navigation, emphasizing performance in various dynamic indoor environments. Furthermore, a multimodal robot navigation model with semantic embeddings utilizing laser and visual input is proposed.

Our future work includes implementing the proposed multimodal robot navigation model and evaluating its navigation performance across various indoor environments using the proposed benchmarking metrics.

Acknowledgement. The Malaysian Ministry of Higher Education supported this research work through the Fundamental Research Grant Scheme (FRGS) under Grant FRGS/1/2022/ICT02/SWIN/02/1.

References

1. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems **2017-December**, 5999–6009 (2017)

2. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3D object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2886–2897 (2021). https://doi.org/10.1109/ICCV48922.2021.00290
3. He, C., Li, R., Li, S., Zhang, L.: Voxel set transformer: a set-to-set approach to 3D object detection from point clouds. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2022-June, pp. 8407–8417 (2022). https://doi.org/10.1109/CVPR52688.2022.00823
4. Mao, J., et al.: Voxel transformer for 3D object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3144–3153 (2021). https://doi.org/10.1109/ICCV48922.2021.00315
5. Saavedra-Ruiz, M., Morin, S., Paull, L.: Monocular robot navigation with self-supervised pretrained vision transformers. In: Proceedings – 2022 19th Conference on Robots and Vision, CRV 2022, pp. 197–20 (2022). https://doi.org/10.1109/CRV55824.2022.00033
6. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: ScanQA: 3D question answering for spatial scene understanding. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19129–19139 (2022). Accessed: 15 Sep 2024. https://openaccess.thecvf.com/content/CVPR2022/html/Azuma_ScanQA_3D_Question_Answering_for_Spatial_Scene_Understanding_CVPR_2022_paper.html
7. Chen, W., Hu, S., Talak, R., Carlone, L.: Leveraging Large (Visual) Language Models for Robot 3D Scene Understanding (2023). arXiv:2209.05629. https://doi.org/10.48550/arXiv.2209.05629
8. Kassab, C., Mattamala, M., Zhang, L., Fallon, M.: Language-extended indoor SLAM (LEXIS): a versatile system for real-time visual scene understanding. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 15988–15994 (2024). https://doi.org/10.1109/ICRA57147.2024.10610341
9. Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. https://arxiv.org/abs/2103.00020v1. Accessed: 06 Sep 2024
10. Chen, R., et al.: CLIP2Scene: Towards Label-Efficient 3D Scene Understanding by CLIP. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7020–7030 (2023). https://openaccess.thecvf.com/content/CVPR2023/html/Chen_CLIP2Scene_Towards_Label-Efficient_3D_Scene_Understanding_by_CLIP_CVPR_2023_paper.html. Accessed: 15 Sep 2024
11. Wang, Z., et al.: SGEITL: scene graph enhanced image-text learning for visual commonsense reasoning. Proc. AAAI Conf. Artific. Intell. **36**(5), 5914–5922 (2022). https://doi.org/10.1609/aaai.v36i5.20536
12. Lin, T.-Y., et al.: Microsoft COCO: Common Objects in Context. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8693 LNCS, no. PART 5, pp. 740–755 (2014)
13. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7
14. Fan, J., Zheng, P., Li, S.: Vision-based holistic scene understanding towards proactive human–robot collaboration. Robot. Comput.-Integr. Manuf. **75**, 102304 (2022). https://doi.org/10.1016/j.rcim.2021.102304
15. Desai, K., Johnson, J.: VirTex: learning visual representations from textual annotations. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11162–11173 (2021). https://openaccess.thecvf.com/content/CVPR2021/html/Desai_VirTex_Learning_Visual_Representations_From_Textual_Annotations_CVPR_2021_paper.html. Accessed: 13 Sep 2024
16. Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., Beyer, L.: Image captioners are scalable vision learners too. https://arxiv.org/abs/2306.07915v5. Accessed: 6 Sep 2024

17. Roli, A., Jaeger, J., Kauffman, S.A.: How organisms come to know the world: fundamental limits on artificial general intelligence. Front. Ecol. Evol. (2022). https://doi.org/10.3389/fevo.2021.806283
18. Peng, S., Genova, K., "Max" Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: OpenScene: 3D Scene Understanding with Open Vocabularies (2023). arXiv:2211.15654. https://doi.org/10.48550/arXiv.2211.15654
19. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: minkowski convolutional neural networks. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3075–3084 (2019). https://openaccess.thecvf.com/content_CVPR_2019/html/Choy_4D_Spatio-Temporal_ConvNets_Minkowski_Convolutional_Neural_Networks_CVPR_2019_paper.html. Accessed 13 Sep 2024
20. Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y.: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels (2022). arXiv:2112.12143. https://doi.org/10.48550/arXiv.2112.12143
21. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven Semantic Segmentation. arXiv:2201.03546 (2022). https://doi.org/10.48550/arXiv.2201.03546
22. Graf, F., Lindermayr, J., Odabaşi, Ç., Huber, M.F.: Toward holistic scene understanding: a transfer of human scene perception to mobile robots. IEEE Robot. Autom. Mag. **29**(4), 36–49 (2022). https://doi.org/10.1109/MRA.2022.3210587
23. Chen, W.: Applications of Large Language Models for Robot Navigation and Scene Understanding. Thesis, Massachusetts Institute of Technology (2023). https://dspace.mit.edu/handle/1721.1/151450. Accessed 15 Sep 2024
24. Sridharan, M., Mota, T.: Combining Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning in Robotics (2022). arXiv:2201.10266. https://doi.org/10.48550/arXiv.2201.10266
25. Song, X., Liang, X., Huaidong, Z.: Semantic mapping techniques for indoor mobile robots: review and prospect. Measur. Control **58**(3), 377–393 (2024). https://doi.org/10.1177/00202940241259903
26. Tang, Y., et al.: Mobile Robot Oriented Large-Scale Indoor Dataset for Dynamic Scene Understanding (2024). arXiv:2406.19791. https://doi.org/10.48550/arXiv.2406.19791
27. Möller, R., Furnari, A., Battiato, S., Härmä, A., Farinella, G.M.: A survey on human-aware robot navigation. Robot. Auton. Syst. **145**, 103837 (2021). https://doi.org/10.1016/j.robot.2021.103837
28. De Magistris, G., et al.: Vision-based holistic scene understanding for context-aware human-robot interaction. In: Bandini, S., Gasparini, F., Mascardi, V., Palmonari, M., Vizzari, G. (eds.) AIxIA 2021 – Advances in Artificial Intelligence: 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1–3, 2021, Revised Selected Papers, pp. 310–325. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-08421-8_21
29. Loganathan, A., Ahmad, N.S.: A systematic review on recent advances in autonomous mobile robot navigation. Eng. Sci. Technol. An Int. J. **40**, 101343 (2023). https://doi.org/10.1016/j.jestch.2023.101343
30. Marashian, A., Razminia, A.: Mobile robot's path-planning and path-tracking in static and dynamic environments: dynamic programming approach. Robot. Auton. Syst. **172**, 104592 (2024). https://doi.org/10.1016/j.robot.2023.104592
31. Scott, K., Foote, T.: "2023 ROS Metrics Report"
32. "Nav2 — Nav2 1.0.0 documentation." https://docs.nav2.org/. Accessed: 14 Sep 2024
33. Macenski, S., Moore, T., Lu, D.V., Merzlyakov, A., Ferguson, M.: From the desks of ROS maintainers: a survey of modern & capable mobile robotics algorithms in the robot operating system 2. Robot. Auton. Syst. **168**, 104493 (2023). https://doi.org/10.1016/j.robot.2023.104493

34. Yin, J., Li, A., Li, T., Yu, W., Zou, D.: M2DGR: A Multi-sensor and Multi-scenario SLAM Dataset for Ground Robots (2021). arXiv:2112.13659. https://doi.org/10.48550/arXiv.2112.13659
35. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and System, pp. 573–580 (2012). https://doi.org/10.1109/IROS.2012.6385773
36. Grupp, M.: *MichaelGrupp/evo*. Python (2024). https://github.com/MichaelGrupp/evo. Accessed 15 Sep 2024
37. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: an accurate open-source library for visual, visual-inertial and multi-map SLAM. IEEE Trans. Robot. **37**(6), 1874–1890 (2021). https://doi.org/10.1109/TRO.2021.3075644
38. Zhang, J., Singh, S.: LOAM: lidar odometry and mapping in real-time. In: Robotics: Science and Systems X, Robotics: Science and Systems Foundation (2014). https://doi.org/10.15607/RSS.2014.X.007
39. Shan, T., Englot, B.: LeGO-LOAM: lightweight and ground-optimized lidar odometry and mapping on variable terrain. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4758–4765 (2018). https://doi.org/10.1109/IROS.2018.8594299
40. Qin, C., Ye, H., Pranata, C.E., Han, J., Zhang, S., Liu, M.: LINS: a lidar-inertial state estimator for robust and efficient navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 8899–8906 (2020). https://doi.org/10.1109/ICRA40945.2020.9197567
41. Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D.: LIO-SAM: tightly-coupled lidar inertial odometry via smoothing and mapping. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5135–5142 (2020). https://doi.org/10.1109/IROS45743.2020.9341176
42. Janoch, A., et al.: A category-Level 3D object dataset: putting the kinect to work. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision: Research Topics and Applications, pp. 141–165. Springer London, London (2013). https://doi.org/10.1007/978-1-4471-4640-7_8
43. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V, pp. 746–760. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
44. Xiao, J., Owens, A., Torralba, A.: SUN3D: a database of big spaces reconstructed using SfM and object labels. In: 2013 IEEE International Conference on Computer Vision, Sydney, Australia: IEEE, pp. 1625–1632 (2013). https://doi.org/10.1109/ICCV.2013.458
45. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2D-3D-Semantic Data for Indoor Scene Understanding (2017). arXiv:1702.01105. https://doi.org/10.48550/arXiv.1702.01105
46. McCormac, J. , Handa, A., Leutenegger, S., Davison, A.J.: SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth (2017). arXiv:1612.05079. https://doi.org/10.48550/arXiv.1612.05079
47. Song, S., Lichtenberg, S.P., Xiao J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 07–12-June-2015, pp. 567–576 (2015). https://doi.org/10.1109/CVPR.2015.7298655
48. Dai, A., et al.: ScanNet: richly-annotated 3D Reconstructions of Indoor Scenes. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 2432–2443 (2017). https://doi.org/10.1109/CVPR.2017.261

49. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic Scene Completion from a Single Depth Image (2016). arXiv:1611.08974. https://doi.org/10.48550/arXiv.1611.08974
50. Chang, A., et al.: Matterport3D: Learning from RGB-D Data in Indoor Environments (2017). arXiv:1709.06158. https://doi.org/10.48550/arXiv.1709.06158
51. Li, W., et al.: InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset (2018). arXiv:1809.00716. https://doi.org/10.48550/arXiv.1809.00716
52. Baruch, G., et al.: ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data (2022). arXiv:2111.08897. https://doi.org/10.48550/arXiv.2111.08897
53. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: RIO: 3D Object Instance Re-Localization in Changing Indoor Environments (2019). arXiv:1908.06109. https://doi.org/10.48550/arXiv.1908.06109
54. Rondoni, C., et al.: Navigation benchmarking for autonomous mobile robots in hospital environment. Sci. Rep. **14**(1), 18334 (2024). https://doi.org/10.1038/s41598-024-69040-z
55. 14:00–17:00, "ISO 13482:2014," ISO. Accessed: 14 Sep 2024. https://www.iso.org/standard/53820.html
56. Kästner, L., et al.: Arena-Rosnav 2.0: a development and benchmarking platform for robot navigation in highly dynamic environments. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11257–11264 (2023). https://doi.org/10.1109/IROS55552.2023.10342152
57. Pérez-Higueras, N., Otero, R., Caballero, F., Merino, L.: HuNavSim: A ROS 2 human navigation simulator for benchmarking human-aware robot navigation. IEEE Robot. Autom Letters **8**(11), 7130–7137 (2023). https://doi.org/10.1109/LRA.2023.3316072