Robust Scene Understanding for Mobile Robots Based on Vision and Deep Learning Models

Leticia C. Pereira[a] and Fernando S. Osório[b]
Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo (USP), São Carlos, SP, Brazil

Keywords: Deep Learning, Computer Vision, Mobile Robots, Autonomous Robots.

Abstract: This paper presents the architecture and results of AIVFusion, a real-time perception system designed to generate a rich, multi-layered understanding of an environment from a single monocular camera for autonomous mobile robots. The system is designed to fuse information from different deep learning models to achieve a comprehensive scene understanding. Our architecture integrates three open-source models to perform distinct perception tasks: object detection (YOLOv8), semantic segmentation (FastSAM), and monocular depth estimation (Depth Anything V2). By fusing these outputs, the system generates a unified representation that identifies the navigable area, detects nearby obstacles based on depth information, and semantically labels those identified as "person". The resulting perceptual information can then be leveraged by higher-level systems for tasks such as decision-making and safer navigation. The system's viability is demonstrated through qualitative tests in indoor environments. These results confirm its ability to operate in real-time (approximately 10 FPS) and to effectively fuse the perception layers, even in challenging scenarios involving partial object occlusion.

1 INTRODUCTION

Autonomous mobile robots (AMRs) have grown rapidly, driven by advances in artificial intelligence, robotics, and increasingly accessible hardware. Their adoption has expanded across sectors due to rising interest from individuals and companies. AMRs are now widely used in applications such as domestic assistance, delivery, warehousing, and logistics, proving to be efficient and innovative solutions. These robots navigate autonomously in both indoor and outdoor environments, adapting to static and dynamic obstacles like moving people (Liu et al., 2024; Niloy et al., 2021). However, this work focuses specifically on indoor scenarios.

Indoor environments present a significant challenge due to their unpredictability, often characterized by high flows of people (dynamic obstacles), and static obstacles that can be moved around the environment. In addition, narrow corridors and varying lighting conditions make the environment even more complex for robot navigation (Zhang et al., 2024).In particular, the latter, along with physical vibrations when operating on different floor types, are known to influence the performance of deep learning systems (Maruschak et al., 2025). Given this scenario and the growth of AMR applications, safe navigation has become an increasingly important topic.

It is important to highlight that, for a robot of this type to navigate autonomously in an environment, there are several development stages involved, such as perception, mapping, localization, control, navigation, and decision-making, among others. One of the most important stages is the perception layer, which is responsible for understanding the surrounding environment through sensors and making decisions based on the collected information, thereby ensuring safe navigation (Liu et al., 2024).

Robotic perception can be achieved through various sensors, including sonar, LiDAR, and cameras. While sonar systems are often limited by their low spatial resolution, LiDAR provides precise 3D measurements, but at a prohibitive cost for many mobile robot applications. In contrast, camera-based systems offer a compelling trade-off between cost and the richness of the data they provide. Recent advances in deep learning, particularly in monocular depth estimation (Masoumian et al., 2022), have further enhanced the viability of using a single camera to infer three-dimensional scene geometry, making it a cost-

[a] https://orcid.org/0009-0000-1209-5690
[b] https://orcid.org/0000-0002-6620-2794

Pereira, L. C. and Osório, F. S.
Robust Scene Understanding for Mobile Robots Based on Vision and Deep Learning Models.
DOI: 10.5220/0013789100003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 2*, pages 386-393
ISBN: 978-989-758-770-2; ISSN: 2184-2809

effective and powerful sensor for modern perception systems.

Recent advances in deep learning have significantly enhanced computer vision, offering powerful tools for robotic perception. However, individual tasks such as object detection, depth estimation, or semantic segmentation, while powerful, provide an incomplete view of the environment. Achieving robust autonomous navigation with low-cost sensors requires the cohesive fusion of these modalities. To address this challenge, this paper introduces AIVFusion, a real-time perception system that integrates these capabilities to generate a rich, contextual understanding of the scene from a single monocular camera.

The main contribution of this work is a hierarchical fusion architecture. Our system first identifies the navigable area through segmentation to establish a baseline for safety. It then augments this with a map of general obstacles based on their physical proximity, derived from depth estimation. Finally, it applies semantic labels ("person") to detected obstacles, enabling more sophisticated decision-making, such as human-aware navigation.

2 RELATED WORKS

Scene perception for autonomous robots has been transformed by advances in deep neural networks, which now form the basis of many vision systems.In the field of object detection, for instance, the YOLO models have established real-time performance as a viable baseline in robotics (Vijayakumar and Vairavasundaram, 2024). Simultaneously, in monocular depth estimation, large-scale architectures like Depth Anything (Yang et al., 2024a) and lightweight models such as FastDepth (Wofk et al., 2019) have proven effective in inferring scene geometry. In semantic scene understanding, segmentation models, from classic Fully Convolutional Networks (FCNs) (Long et al., 2015) to more recent approaches like Segment Anything (SAM) (Kirillov et al., 2023) and its variants, such as FastSAM (Zhao et al., 2023), enable the semantic classification of pixels into various categories within a scene. While these tools perform well in their individual tasks, integrating information extracted by different modules still represents a challenge for perception. In this section, we review existing fusion strategies to contextualize our proposed approach.

A line of research for collision avoidance focuses on the fusion of object detection with depth estimation. A representative example is the work of (Urban and Caplier, 2021), who propose a system to predict the Time to Collision (TTC). Their architecture first employs an object detector (YOLOv3) to locate predefined classes, such as "Person" and "Chair", followed by a depth estimation network (FastDepth) to estimate the distance to those specific targets. However, this strategy presents two significant limitations. First, its effectiveness is limited to a predefined set of object classes, which may reduce the system's ability to respond to previously unseen or unlabeled obstacles. Second, if an obstacle is missed, for instance, due to partial occlusion, the entire risk assessment process for that object may be compromised.

Another line of research in perception focuses on the direct segmentation of the navigable space. In this context, (Dang and Bui, 2023) demonstrate a technique that uses a binary segmentation network, specifically trained to classify the environment into navigable and non-navigable areas, generating a Bird's-Eye-View (BEV) map for an A* path planner. Although effective in scenarios similar to its training dataset, training a segmentation network on a visually specific dataset presents significant generalization challenges. The approach becomes prone to errors when exposed to new or altered environments, where, for example, changes in lighting conditions or the presence of dynamic obstacles not seen during training can lead to incorrect space segmentation. This, in turn, may cause the planner to generate an erroneous trajectory over an obstacle that was mistakenly classified as a navigable area.

Given these limitations, we argue that a robust perception system capable of supporting autonomous navigation and decision-making layers should integrate multiple sources of environmental information to enable a more holistic perception. Our approach is based on a monocular camera and proposes the fusion of three complementary perception modalities: (1) navigable space segmentation, performed by a fundamental segmentation model (FastSAM), whose large-scale pretraining favors better generalization across different scenarios; (2) explicit depth estimation, through a deep learning model, which enables more robust and class-independent obstacle detection; and (3) semantic object detection, which enables differentiated behaviors, such as social navigation around people.

3 PROPOSED APPROACH

The proposed system AIVFusion is designed to generate a real-time, multi-layered scene understanding from a single monocular camera. Its architecture is based on the parallel execution of three specialized

perception modules, whose outputs are subsequently fused in a hierarchical process. The selection of each module was guided by a balance between state-of-the-art performance and real-time processing capability, as detailed below.

3.1 Architectural Components of the Perception System

The proposed system AIVFusion aims to extract and combine different information from the images, using a monocular camera. The proposed approach considers the fusion of three extracted information layers: image Semantic Segmentation, image Depth Estimation of scene elements, and image Objects Detection and Classification.

Navigable Space Segmentation: The foundation of our perception stack is the identification of the navigable area. For this task, the primary requirement was a model capable of real-time inference. We initially considered a state-of-the-art model like Segment Anything 2 (SAM 2); however, our performance analysis (detailed in Section 4.2) revealed a low framerate of 2 FPS, making it unfeasible for our application. This performance difference stems from their distinct architectures: SAM 2 relies on a computationally intensive Vision Transformer (ViT) as its image encoder (Ravi et al., 2024). In contrast, FastSAM achieves its high speed by leveraging a lightweight CNN-based detector (a YOLOv8-seg model) to generate segmentation proposals (Zhao et al., 2023). Consequently, we adopted FastSAM for its optimal balance of segmentation quality and real-time performance. Both, adopt the task of Prompt-able Segmentation Task, meaning they can segment images using different types of inputs, such as points, bounding boxes, texts, or masks. This flexibility allows the model to segment specific objects in an image based on the provided prompt type. We utilize a point prompt, strategically placed in the lower portion of the camera's view, to robustly isolate the ground plane.

Depth Estimation: To understand the scene's geometry and the proximity of objects, we employ Depth Anything V2. This state-of-the-art model, is a powerful monocular depth estimation model that can be easily deployed to estimate depth in images and also in videos. This model is based on a Vision Transformers (ViT) and zero-shot deep estimation, as in SAM model, that allows the model to handle data and scenes that are not included in its training set (improved generalization and better performance even with unseen scenes) (Yang et al., 2024b). This component is crucial for our class-agnostic obstacle detection, as it allows the system to perceive any physical barrier based solely on its distance, without prior knowledge of its category.

Semantic Object Detection: To add a semantic layer, we selected a model from the You Only Look Once (YOLO) family, a series of real-time object detectors based on deep learning convolutional neural networks (CNNs), known for their efficiency in time-critical applications. Our objective was to use the latest version available. However, this presented a critical dependency conflict: YOLO11 (Khanam and Hussain, 2024) require a version of the ultralytics library (v8.3.0+) that proved incompatible with the version required by FastSAM (v8.0.120). To ensure the stability and integration of the complete pipeline, we selected YOLOv8 (Jocher et al., 2023), as it represents the most capable model within the compatible dependency ecosystem. In our application, it is specifically configured to identify the "person" class.

3.2 The AIVFusion Proposed System

AIVFusion operates by analyzing each video frame from a monocular camera to build a rich understanding of the scene. For each captured frame, the three perception models described above - object detection (YOLOv8), ground segmentation (FastSAM), and depth estimation (Depth Anything V2) - are executed to extract the base information. From this execution, we obtain three primary data layers: (1) the scene's depth map, (2) the navigable area mask, and (3) the location of people (if present). The core of our contribution, detailed below, lies in how these three layers of information are combined to generate a single representation of the environment, highlighting safe areas and potential risks.

As illustrated in the system architecture diagram (Figure 1), the pipeline starts with the capture of the frame by the camera (original input), which in turn is processed in parallel by three deep neural networks, each responsible for extracting a fundamental layer of information from the scene:

- *Depth Estimation:* The first layer generates a dense depth map, a 2D matrix in which each pixel represents the relative distance of a point in the scene relative to the camera. In our implementation, higher values indicate closer proximity. For visualization purposes, this map is converted into a grayscale image, where brighter pixels correspond to nearer objects.
- *Ground Segmentation:* The second layer uses the FastSAM model to isolate the navigable area. Although the model supports text prompts (such as "ground"), it relies on additional language–image

models like CLIP (Radford et al., 2021), which increases processing latency and reduces FPS. For this reason, the point prompt was chosen, as it offers a significantly lower computational cost. A fixed point in the lower central region of the image was defined as a reference, generating a binary mask (ground_mask) that defines the traversable space, visually represented in blue.

- *Person Detection:* The third layer employs the YOLO model for a specific task: finding instances of the "person" class. The output of this module is a set of bounding boxes that define the positions of the detected people in the image.

With the extracted information, the fusion process is initiated. The first stage focuses on the geometry of the environment to identify obstacles in a class-independent manner.

First, a distance threshold is applied to the generated depth map to create the **Initial Collision Mask Calculation**, which filters pixels considered to represent nearby obstacles. Since the depth values are relative, this threshold was empirically defined through controlled experiments: objects were placed at different distances from the camera to identify the approximate pixel values in the depth matrix corresponding to up to 1 meter. Thus, any obstacle detected within this range is considered relevant for immediate risk. The result of this step, visualized in green, is a risk map that, by considering only proximity, initially includes the ground itself.

Next, this information is refined to generate the **Final Collision Mask**. The ground_mask, obtained from segmentation, is logically subtracted from the Initial Collision Mask Calculation. The result is an accurate map that represents only the nearby obstacles that are not part of the navigable area.

The last stage of the pipeline is the fusion with semantic data, where meaning is assigned to the detected obstacles. The bounding boxes of people, detected by YOLO, are overlaid with the Final Collision Mask. When a person is identified within a risk area, the system recognizes not only the presence of an obstacle but also that it is a human, enabling higher-level modules to make more appropriate decisions, such as reducing speed, performing cautious deviations, or even initiating audio interactions.

The output of our system is a rich and contextualized scene understanding, composed of three crucial layers of information: a map of the navigable area, a real-time assessment of nearby obstacles, and the semantic identification of people within risk zones. This unified perception serves as a crucial input for any application requiring spatial and semantic awareness. While particularly relevant for the decision-making and navigation stacks of mobile robots, it also extends to domains where rich contextual scene understanding is required.

Since the objective of our work was to evaluate our fusion architecture's ability to generate a rich, multi-layered understanding of the environment in real-time, rather than to establish a new state-of-the-art in accuracy, we used the original pre-trained weights for all models. The selected models are well-established, with YOLOv8 trained on the COCO dataset, FastSAM efficiently trained on a 2% subset of the large SA-1B dataset, and Depth Anything V2 on a combination of high-quality synthetic data and millions of pseudo-labeled real-world images.

4 RESULTS

4.1 Experimental Setup

The perception system was implemented in Python 3.8.10 using the PyTorch and Ultralytics libraries, among others. All processing was performed on a notebook with the following specifications: an Intel® Core™ i7-7700HQ processor, 16 GB of DDR4 memory, and an NVIDIA GeForce GTX 1050Ti GPU with 4 GB of VRAM, running the Ubuntu 20.04.6 LTS operating system. Images were captured in real time using a Microsoft LifeCam HD-3000 webcam at a resolution of 640×480 pixels. To simulate a real-world application, the camera was mounted on top of the chassis of a mobile robot (Pioneer 3-AT), positioned approximately 30 cm above the ground.

4.2 Real-Time Performance Analysis

The viability of a perception system for mobile robotics is directly related to its ability to operate in real time. To select the models that compose our architecture, a performance analysis was conducted based on the frames per second (FPS) of the main components.

To ensure real-time applicability, it was necessary to optimize the inference speed of the most computationally intensive models. Given the limitations of the available hardware, the strategy adopted was to reduce the input image resolution for the segmentation and depth estimation modules. During inference, the parameters were set to *imgsz=256* for FastSAM and *input_size=256* for Depth Anything v2, as required by each model. This process resizes the original frame so that its shorter side measures 256 pixels while maintaining the aspect ratio, before being fed into the models. Table 1 summarizes the approximate FPS and la-

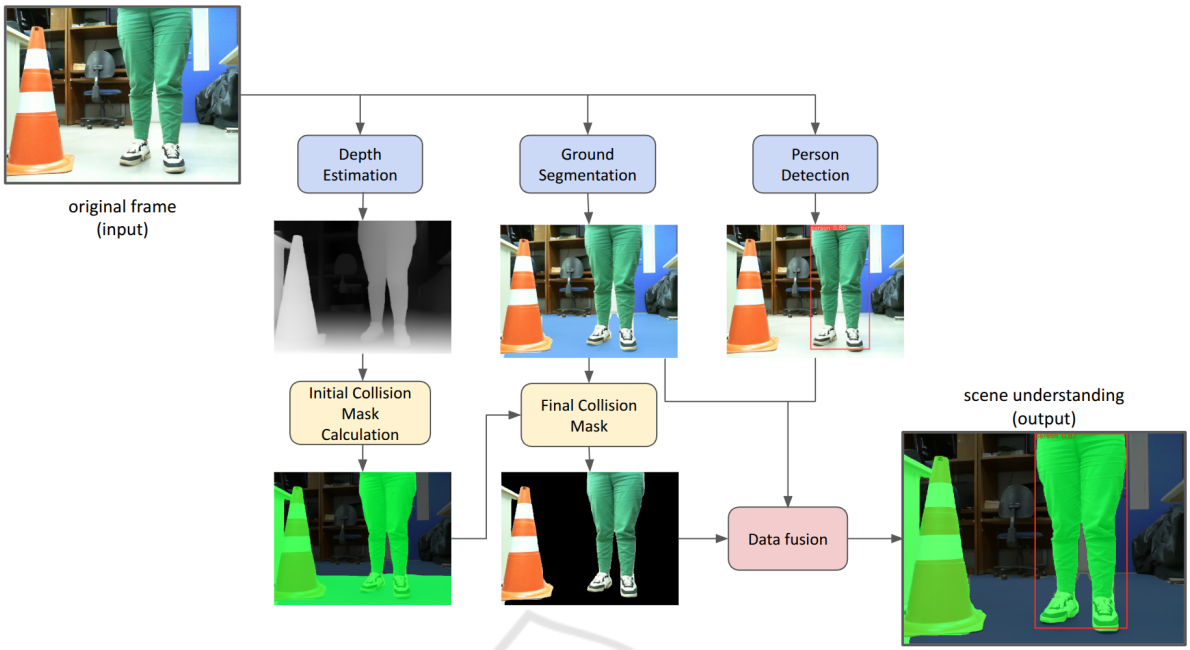

Figure 1: AIVFusion Architecture Diagram (Source: Authors).

tency results for each individual model and, finally, for their integration.

Table 1: Performance Comparison (FPS and Latency) of the Evaluated Models.

Model	Pre-trained Weights	Input Resolution	FPS	Latency(ms)
SAM 2	sam2_hiera_tiny	640 × 480	2	500
FastSAM	FastSAM-s	640 × 480	15–20	50-66
Depth Anything V2	vits	640 × 480	20–24	41.7-50
YOLOv8	yolov8n	640 × 480	25–30	33-40
Integrated System	-	640 × 480	10	100

The analysis in Table 1 shows that, although SAM 2 is a powerful model, its performance of approximately 2 FPS makes it unfeasible for our application. In contrast, the models YOLOv8, FastSAM, and Depth Anything V2, running with optimized input resolutions and smaller pre-trained weights, achieved suitable frame rates individually. When integrated into our fusion pipeline, the system reached 10 FPS on average. It is important to highlight that this performance was only possible after optimizing the visualization routines, particularly by excluding FastSAM's *plot_to_result()* function, which had significantly reduced the FPS. However, this function is intended solely for visualization purposes and is not required in the actual application, making its removal feasible.

This resulting performance of 10 FPS is a key outcome for the practical application of our system. It translates to a total perception pipeline latency of approximately 100 ms (Tab. 1). This response time is well within the soft real-time requirements for safe navigation of a mobile robot at low to medium speeds, as it allows the sense-plan-act cycle to update the world model and react to obstacles effectively. Therefore, this result validates the viability of our fusion architecture, demonstrating that it is possible to combine the rich, multi-modal capabilities of these models while maintaining the real-time performance essential for autonomous robotic tasks.

4.3 Qualitative Analysis of Perception and Fusion

To validate the effectiveness and generalization capabilities of the AIVFusion architecture, qualitative tests were conducted across three distinct scenarios: a controlled laboratory environment, a complex indoor common area, and a challenging outdoor environment.

The initial validation was performed in our laboratory setting to confirm the core functionality of the fusion pipeline 2. Figure 2(a) presents the original scene captured by the camera. From this input, the system first identifies the traversable space, as shown in Figure 2(b), where the ground segmentation mask (in blue) defines the navigable area. Next, the fusion logic is applied to generate the final collision mask, shown in Figure 2(c). At this stage, the system isolates only the objects that represent a real proximity risk, with the ground plane already excluded. The

unified output of the system is shown in Figure 2(d), where all information layers are overlaid: the obstacle (in green) is contextualized with its semantic label ("person"). It is worth noting that the person was successfully detected despite significant partial occlusion, demonstrating the system's capability to generate coherent scene understanding in challenging situations.

To further assess the system's performance in a new indoor environment, tests were conducted in a corporate common area (Figure 3). While the fusion of ground segmentation and depth-based obstacles remained robust, this scenario highlighted the limitations of the underlying perception models. In one case (Figure 3, top row), the object detector produced a false positive, identifying three people where only two were present. In another instance (Figure 3, bottom row), the collision mask only partially covered the person, suggesting that the empirically set depth threshold may require fine-tuning or dynamic adaptation for different scenes. These observations are valuable as they demonstrate the overall architecture's resilience while pinpointing areas for future improvements in the individual perception components.

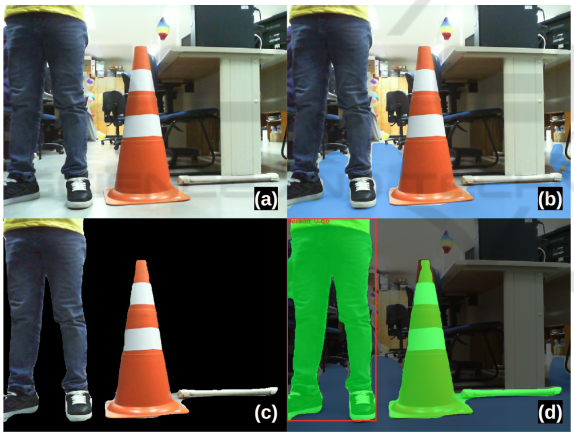

Figure 2: The Visual Perception Pipeline for Scene Understanding (Source: Authors).

Finally, to test the system's generalization limits, a preliminary evaluation was conducted in an outdoor residential environment (Figure 4). The segmentation module successfully recognized the sidewalk as a navigable surface and, importantly, identified the adjacent lawn as an unsafe area (obstacle), a correct and crucial inference for ensuring robot safety. The full pipeline also detected a person on the sidewalk as a potential risk, showing promising potential for generalization beyond structured indoor environments.

While our qualitative results are strong, our architectural design also considered the previously mentioned real-world challenges, such as varying lighting

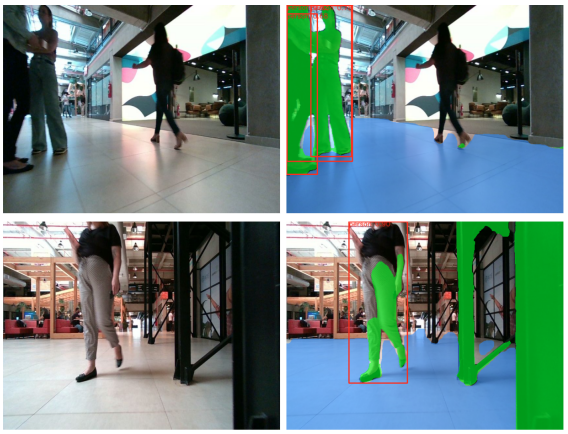

Figure 3: Qualitative Results in a Novel Indoor Environment (Source: Authors).

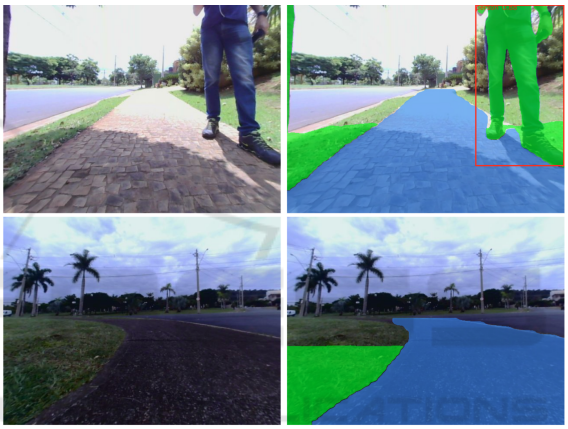

Figure 4: Preliminary Evaluation of System Generalization in an Outdoor Scenario (Source: Authors).

and vibrations from different floor types. Our system addresses this by balancing the capabilities of its component models: we balance the strong domain generalization of the ViT-based Depth Anything V2, which is crucial for adapting to visual variations caused by environmental changes (Alijani et al., 2024), with the high-speed performance of our selected CNN-based models, FastSAM and YOLOv8. This combination allows the system to produce a rich perceptual output while maintaining real-time performance.

5 CONCLUSIONS AND FUTURE WORKS

This paper introduced and qualitatively validated AIVFusion, a hierarchical perception architecture that fuses object detection, semantic segmentation, and monocular depth estimation from a single camera. This work's primary contribution is the demonstra-

tion that such a complex fusion is viable for real-time applications, achieving approximately 10 FPS on consumer-grade hardware while robustly identifying navigable areas and obstacles. Qualitative results confirmed the effectiveness of the proposed fusion logic, showing that the system robustly identifies navigable areas and obstacles in varied indoor environments. Furthermore, preliminary outdoor tests indicated promising generalization capabilities, where a sidewalk was correctly interpreted as navigable terrain. The system also proved resilient in challenging scenarios, such as those involving significant partial object occlusion. These findings validate that the proposed approach provides a comprehensive scene understanding.

As future work, the next crucial step is rigorous quantitative validation of the perception system. To this end, a custom and diverse dataset is currently being developed. This dataset will consist of video sequences captured in multiple indoor scenarios, under various lighting conditions, and will include a range of static and dynamic obstacles to test the system's limits. It will contain annotated ground truth for each of the system's outputs - navigable space, obstacles, and people, which will allow an objective evaluation of performance. The evaluation metrics will be selected based on standard practices in the literature for each perception task. The goal of this quantitative validation is therefore to objectively assess the effectiveness of the proposed fusion architecture and to test the hypothesis that it contributes to more accurate scene understanding. Beyond this validation, future works aim to evaluate the system in more complex scenarios and to integrate the proposed perception layer into a navigation pipeline for autonomous robot operation. We also consider improving the system performance using multiple Processor Cores and NPUs (e.g., Intel Core Ultra 9 Hardware), and GPUs (Nvidia Jetson and Nvidia Digits).

DATA AVAILABILITY STATEMENT

The source code developed for this research is not yet publicly available, but can be provided by the corresponding author upon reasonable request. The code is currently being prepared for public release and will be made available in a dedicated repository.

ACKNOWLEDGEMENTS

This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. The authors also wish to thank the Institute of Mathematical and Computer Sciences (ICMC/USP) and the Center of Excellence in Artificial Intelligence (CEIA) for their support.

REFERENCES

Alijani, S., Fayyad, J., and Najjaran, H. (2024). Vision transformers in domain adaptation and domain generalization: a study of robustness. *Neural Computing and Applications*, 36(29):17979–18007.

Dang, T.-V. and Bui, N.-T. (2023). Obstacle avoidance strategy for mobile robot based on monocular camera. *Electronics*, 12(8):1932.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8.

Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. https://arxiv.org/abs/2410.17725.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything. https://arxiv.org/abs/2304.02643.

Liu, Y., Wang, S., Xie, Y., Xiong, T., and Wu, M. (2024). A review of sensing technologies for indoor autonomous mobile robots. *Sensors*, 24(4).

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Maruschak, P., Konovalenko, I., Osadtsa, Y., Medvid, V., Shovkun, O., and Baran, D. (2025). Surface defects of rolled metal products recognised by a deep neural network under different illuminance levels and low-amplitude vibration. *The International Journal of Advanced Manufacturing Technology*, pages 1–16.

Masoumian, A., Rashwan, H. A., Cristiano, J., Asif, M. S., and Puig, D. (2022). Monocular depth estimation using deep learning: A review. *Sensors*, 22(14):5353.

Niloy, M. A. K., Shama, A., Chakrabortty, R. K., Ryan, M. J., Badal, F. R., Tasneem, Z., Ahamed, M. H., Moyeen, S. I., Das, S. K., Ali, M. F., Islam, M. R., and Saha, D. K. (2021). Critical design and control issues of indoor autonomous mobile robots: A review. *IEEE Access*, 9:35338–35370.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Urban, D. and Caplier, A. (2021). Time-and resource-efficient time-to-collision forecasting for indoor pedestrian obstacles avoidance. *Journal of imaging*, 7(4):61.

Vijayakumar, A. and Vairavasundaram, S. (2024). Yolo-based object detection models: A review and its applications. *Multimedia Tools and Applications*, 83(35):83535–83574.

Wofk, D., Ma, F., Yang, T.-J., Karaman, S., and Sze, V. (2019). Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. (2024a). Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. (2024b). Depth anything v2. https://arxiv.org/abs/2406.09414.

Zhang, Y., Liu, Y., Liu, S., Liang, W., Wang, C., and Wang, K. (2024). Multimodal perception for indoor mobile robotics navigation and safe manipulation. *IEEE Transactions on Cognitive and Developmental Systems*.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., and Wang, J. (2023). Fast segment anything. *arXiv preprint arXiv:2306.12156*.