Indoor Semantic Scene Understanding Using Sensor Fusion Approach

Dissertation submitted in partial fulfilment of the requirement for the degree of

Bachelor of Technology in Computer Science and Engineering

Submitted by

Deep senchowa (CSE-09/20)
Deepjyoti Kalita (CSE-10/20)
Priyakshi Bordoloi (CSE-27/20)

Under the guidance of
Dr. Abhijit Boruah
Assistant Professor
Department of Computer Science and Engineering
DUIET, Dibrugarh University

Department of Computer Science and Engineering
Dibrugarh University Institute of Engineering and Technology
Dibrugarh University
Dibrugarh-786004, Assam

CERTIFICATE

This is to certify that the report entitled "**Indoor Semantic Scene Understanding Using Sensor Fusion Approach**" is a Bonafede work carried out by Deep Senchowa (CSE-09/20), Deepjyoti Kalita (CSE-10/20), and Priyaksi Bordoloi (CSE-27/20) in partial fulfilment for the award of the degree of Bachelor of Technology in Computer Science and Engineering from Dibrugarh University Institute of Engineering and Technology under Dibrugarh University, Dibrugarh, Assam during the academic year 2023-2024. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The Project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Technology degree.

Dr. Abhijit Boruah
Project Supervisor
Dept. of Computer Science and Engineering
DUIET, Dibrugarh University

Forwarded by:

Head of the Department
Dept. of Computer Science and Engineering
DUIET, Dibrugarh University Date:

Examiners

(Internals) (Externals)

DECLARATION

We do hereby declare that the work presented in this dissertation is exclusively our own under the supervision of Dr. Abhijit Boruah, Assistant Professor, Department of Computer Science and Engineering, Dibrugarh University Institute of Engineering and Technology, Dibrugarh University and was carried out by the authors. The extent and source of information derived from existing literature have been indicated in the dissertation at appropriate places.

Date:
Place: DUIET, Dibrugarh

Deep Senchowa (CSE-09/20)

Deepjyoti Kalita (CSE-10/20)

Priyakshi Bordoloi (CSE-27/20)

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our project supervisor Dr. Abhijit Boruah for his guidance, invaluable support and encouragement throughout the semester. This project would have been impossible without our project supervisor and we want to extend our sincere gratitude to him for his guidance and constant supervision as well as for providing necessary information regarding the project.

We would also like to express our gratitude towards the teachers and staff of the Department of Computer Science and Engineering of Dibrugarh University Institute of Engineering and Technology, Dibrugarh University for their kind co-operation and encouragement, which helped us in completing this project. Our thanks and appreciation goes to our batch-mates in developing the project and people who have directly or indirectly helped us out with their abilities.

Above all, we would like to thank our parents and family members for their support all along.

Deep Senchowa (CSE-09/20)

Deepjyoti Kalita (CSE-10/20)

Priyakshi Bordoloi (CSE-27/20)

ABSTRACT

This project aims to advance indoor scene understanding through the synergistic fusion of 2D Light Detection and Ranging (LiDAR) and camera sensor modalities. The combination of these sensors provides a robust and comprehensive perception system, addressing the limitations of individual sensor technologies in complex indoor environments. The project leverages the strengths of 2D LiDAR for precise distance measurements and camera sensors for high-resolution imaging, thereby enhancing the overall perception accuracy. We trained the models with our own tabletop dataset. We are doing semantic segmentation using CNN model for object detection to achieve higher accuracy to its competitors.

Keywords: Image segmentation, Image classification, YOLO V8, semantic segmentation

Table of Contents

CERTIFICATE i

DECLARATION ii

ACKNOWLEDGEMENT iii

ABSTRACT iv

INTRODUCTION 1

RELATED WORK 3

METHODLOGY 6

3.1 Joint Calibration 6

3.2. Features Extraction 7

3.3. Object Detection and Segmentation 8

3.4 Semantic Segmentation 9

3.5 YOLO-V8 10

Implementation 11

References 14

Chapter 1

INTRODUCTION

In the realm of computer vision and robotics, the profound comprehension of indoor environments plays a pivotal role in enhancing the capabilities of autonomous systems. The synergy of camera and 2D LiDAR (Light Detection and Ranging) sensors has emerged as a potent solution for addressing the challenges associated with indoor scene understanding. This project delves into the intricacies of fusing visual and depth perception to decipher the spatial relationships between objects within a confined space.

The fundamental goal of indoor scene understanding lies in empowering machines to interpret the complex interplay of objects within their surroundings. Beyond mere object recognition, the focus here is on elucidating the spatial context that governs the relative positioning of objects. Identifying which object is in front of, behind, to the left or right, or even atop another, represents a significant leap towards endowing machines with a more nuanced comprehension of their environment.

The utilization of cameras provides rich visual information, enabling the system to recognize objects based on their appearance and structure. On the other hand, 2D LiDAR complements this visual data by offering precise depth information, facilitating the creation of accurate spatial maps. The amalgamation of these two sensor modalities establishes a robust foundation for our endeavour to unravel the spatial relationships within an indoor setting.

As technology advances, the applications of indoor scene understanding extend across diverse domains, including robotics, smart environments, and augmented reality. This project aligns with the broader objective of advancing the field by contributing innovative solutions to the intricate challenge of comprehending indoor spaces through the fusion of camera and 2D LiDAR data. Through this exploration, we aim to enhance the perceptual capabilities of autonomous systems, thereby fostering a more seamless integration of machines into human-centric environments.

The objective of our study is to:

- Develop a method to analyses the spatial arrangement of objects within indoor scenes, leveraging the combined information from cameras and 2D LiDAR sensors.
- Integrate depth perception derived from LiDAR data to enhance the accuracy of object relationships, particularly in scenarios where visual information alone may be ambiguous.
- Infer semantic relationships, such as objects being in close proximity, adjacent, in front of, or on top of one another, fostering a richer understanding of the indoor environment.

Chapter 2

RELATED WORK

A comprehensive review of existing literature reveals a burgeoning interest in indoor scene understanding, with researchers employing a variety of techniques to tackle the intricacies of spatial relationships between objects. Notably, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for visual perception tasks. Our paper review encompasses an exploration of different CNN models applied to indoor scene understanding, aiming to distil insights from state-of-the-art approaches.

The works surveyed span diverse methodologies, ranging from traditional image-based CNNs to more sophisticated models that integrate depth information from LiDAR sensors. Notable studies include [1] which introduced a novel CNN architecture tailored for indoor scene understanding, and [2],which investigated the fusion of camera and LiDAR data to enhance the accuracy of object localization and spatial relationships.

By synthesizing knowledge from these seminal works, we aim to build upon the current understanding of indoor scene understanding and contribute novel insights that advance the state-of-the-art in this burgeoning field. Through this project, we aspire to not only decipher the intricate relationships between objects but also to refine the applicability of existing CNN models in the context of indoor spatial perception.

A summary of some indoor scene understanding paper review is shown in the table 1.1

Paper Title	Author & Year	Hardware	Algorithm/Approaches	Environment	Limitations
Self-Learning Exploration and Mapping for Mobile Robots via Deep Reinforcement Learning. [3]	Fanfei Chen, Shi Bai, Tixiao Shan, Brendan Englot. (2019)	Range sensor (Lidar)	Deep Reinforcement Learning, RNN, DRL	Indoor	1. Limited quantitative data on performance. 2. Suboptimal performance if training data is not representative. 3. Data representativeness.
Object Semantic Grid Mapping with 2D LiDAR and RGB-D Camera for Domestic Robot Navigation. [4]	Xianyu Qi, Wei Wang, Ziwei Liao, Xiaoyu Zhang, Dongsheng Yang, Ran Wei. (2020)	2D LiDAR, RGB-D camera.	SLAM, object-oriented minimum bounding rectangles (OOMBRs).	Indoor	1. Complexity in system integration. 2. Spiral visual maps may limit certain path planning tasks. 3. Limited sensor fusion.
3D Scene Reconstruction Based on a 2D Moving LiDAR. [5]	Harold F. Murcia(B), Maria Fernanda Monroy, Luis Fernando Mora (2018)	2D laser scanner (Hokuyo URG-04LX-UG01), Step motor.	Experimental acquisition algorithm implemented ROS in Python language, MATLAB	Indoor	Shiny objects or elements generate measurement error.
A mobile Robot Mapping Method Integrating LiDAR and Depth Camera [6]	Lu Jia, Zhengguang Ma, Yongguo Zhao (2022)	Depth camera, LiDAR	RTAB-MAP, ROS-based mobile robot platform.	Indoor	1. LiDAR cannot collect information that are not in the horizontal range. 2. Lead to the possibility of collision with obstacles.
2D LiDAR and Camera Fusion in 3D Modelling of Indoor Environment. [7]	Juan Li, Xiang He, Jia Li. (2015)	2D line-scan LiDAR, Digital camera	Joint collaboration, semantic mapping, data association	Indoor	1. Fragmented work in indoor semantic mapping. 2. Relying on a single cue for scene estimation. 3. Inaccurate edge mapping.

					4. Lack of coupling model between semantic categories.
A survey of LiDAR and camera fusion enhancement. [8]	Huazan Zhonga , Hao Wangb , Zhengrong Wua , Chen Zhangc , Yongwei Zhengc , Tao Tangd, (2020)	LiDAR camera fusion	ResNet, CSPN, CSPN++, GuideNet, AVOD, SCA, ESU	Indoor	1. The distribution and types of features in the environment are uneven, and large differences exist. 2. which can decrease the performance of the perception system. 3. Large differences in performance evaluation
LiDAR and Camera Fusion Approach for Object Distance Estimation in Self-Driving Vehicles [9]	G Ajay Kumar, Jin Hee Lee, Jongrak Hwang, Jaehyeong Park, Sung Hoon Youn, Soon Kwon (2020)	Camera, LiDAR, RefineDet detector.	Depth estimation using sensor fusion data, RANSAC plane segmentation, e RANSAC sphere detection, iterative closest point (ICP)	Indoor	Average error rate of each radius level increases as the distance increases.
Extrinsic Calibration of a Camera and Laser Range Finder (improves camera calibration) [10]	Qilong Zhang, Robert Pless (2015)	Camera, 2D laser range finder			

Table:1.1 (some related paper review and compression table)

Chapter 3

METHODLGY

Object-augmented semantic scene classification refers to a scene understanding approach that incorporates detailed information about individual objects in the environment as a part of the semantic segmentation process. It goes a step further than the traditional semantic scene understanding by recognizing and incorporating specific instances of objects into the semantic map [3].

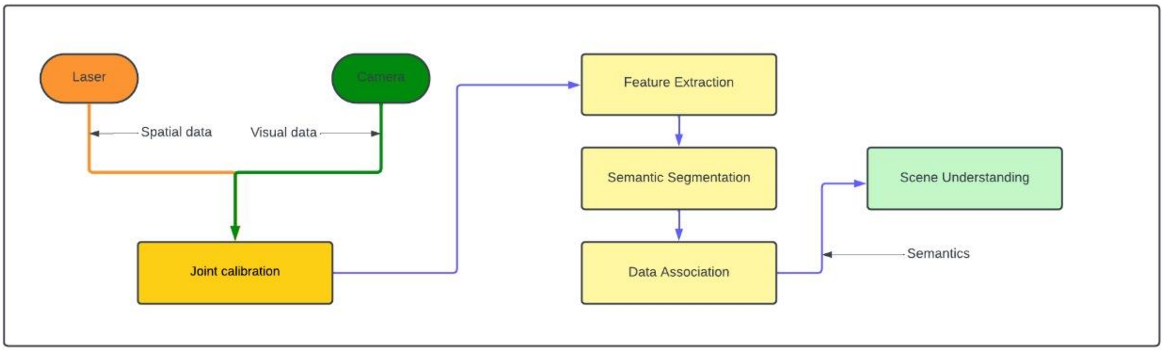


Fig: 1.1 Model architecture

3.1 Joint Calibration

In this project, we use the centre point of the bounding box to denote the pixel coordinates of objects in images. After obtaining the detection semantic information of objects, it needs to be mapped into the map based on the relative relationship between the pixel coordinate frame and the laser coordinate frame. Since the visual information of the camera and the laser data are not consistent, data alignment between the camera and the laser is essential by means of joint calibration, as shown in Fig.2.1.

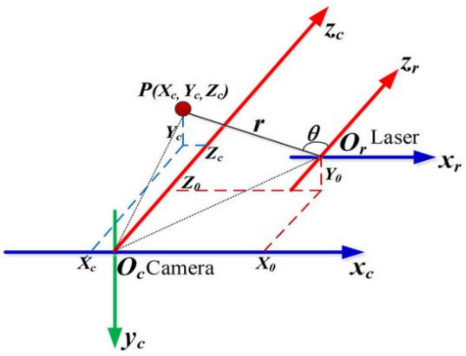


Fig:2.1

The coordinates of the detection object P in laser coordinate frame, camera coordinate frame and pixel coordinate frame are (r, θ), (Xc, Yc, Zc), and (u, v), respectively. Then there are:

$$\begin{pmatrix} Xc \\ Yc \\ Zc \end{pmatrix} = \begin{pmatrix} Xo \ + r\sin\theta \\ Yo \\ Zo + r\ cos\theta \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} u \\ v \\ l \end{pmatrix} = \begin{pmatrix} fx * \frac{X_0 + r\ sin\theta}{Z_0 + r\ cos\theta} \ + \ C_x \\ f_y * \frac{X_0 + r\ sin\theta}{Z_0 + r\ cos\theta} \ + \ C_y \\ 1 \end{pmatrix} \tag{2}$$

where (X0, Y0, Z0) is the coordinates of the laser in the camera coordinate system, fx , fy , cx , cy are internal parameters of camera. The above formula is a typical nonlinear equation. The joint calibration process of the laser and the camera is to solve the above formula. To realize data fusion, we adopt the method in work [3] to perform joint calibration to obtain the coordinate transformation relationship between the camera and the laser.

3.2. Features Extraction

Spatial feature extraction from 2D LiDAR and camera data is a critical process in achieving a comprehensive understanding of indoor scenes. The 2D LiDAR data, processed into an occupancy grid map, provides essential spatial information about the environment, including object locations and distances. Simultaneously, camera data offers rich visual details and semantic information, allowing for the identification of objects and their boundaries. Integrating these two sensor modalities involves aligning their respective coordinate frames and fusing their extracted features synergistically. The combined features provide a holistic representation of the indoor space, enhancing scene understanding capabilities. This integrated approach leverages the strengths of LiDAR and camera sensors, offering a more robust and nuanced perception system that is essential for applications such as robotics, autonomous navigation, and smart environments.

3.3. Object Detection and Segmentation

Our approach to object detection and segmentation is grounded in the utilization of YOLO-V8 (You Only Look Once, Version 8), a state-of-the-art deep learning architecture renowned for its real-time and high-accuracy performance in object detection tasks. The methodology can be delineated into two key stages: object detection and subsequent instance segmentation.

- **Object Detection and Segmentation using YOLO-V8:**

Our object detection pipeline employs the YOLO-V8 model to discern and localize objects within the given indoor scene. YOLO-V8 operates on the principle of dividing the image into a grid and predicting bounding boxes and class probabilities for each grid cell. This enables the model to make predictions at an impressive speed, making it well-suited for real-time applications.

The model is pretrained on a diverse dataset to recognize a wide array of objects, ensuring its adaptability to the indoor environment under consideration. During the inference phase, the YOLO-V8 model processes the input image through a deep neural network and outputs bounding boxes encompassing detected objects, along with corresponding class probabilities.

Following object detection, we employ a segmentation module to refine the understanding of object boundaries, facilitating a more detailed delineation of each object in the scene. We utilize YOLO-V8's inherent ability to predict object masks, which correspond to the pixel-wise segmentation of individual objects.

The segmentation is achieved by extracting pixel-level information from the output feature maps of YOLO-V8. Each detected object is associated with a mask, representing the specific pixels belonging to that object. This segmentation step not only enhances the precision of object localization but also enables a more granular understanding of the spatial extent of each object.

The combination of object detection and segmentation using YOLO-V8 results in a comprehensive representation of the indoor scene, providing both bounding box coordinates and pixel-level segmentation masks for each detected object. This rich information forms the basis for subsequent analyses and applications, including spatial relationship understanding between objects.

3.4 Semantic Segmentation

In this project, semantic segmentation is performed by fusing features from a 2D lidar and RGB camera. The depth information from the lidar provides valuable spatial cues that complement the visual data from the camera.

The lidar depth map is aligned and calibrated to the camera image so that each pixel correspondence is known. The aligned camera image and lidar depth map are passed through an encoder-decoder neural network to extract fused features capturing information from both modalities.

These fused features encode knowledge of appearance from the RGB data and shape/structure from the depth data. They serve as strong representations for differentiating between semantic classes in the scene.

The fused features are passed to a semantic segmentation network based on a FCN, U-Net or other architecture [5]. This network is trained to classify each pixel in the image/depth map into semantic categories like road, sidewalk, vehicles, buildings etc.

By leveraging complementary data from lidar and camera, the model is able to achieve more accurate and robust semantic segmentation compared to using either sensor alone. The lidar depth provides important spatial context while the camera gives detailed appearance information.

In summary, semantic segmentation is performed by an end-to-end model that encodes lidar and camera data into a joint representation to differentiate pixel-level semantic classes in the scene.

3.5 YOLO-V8

YOLOv8 is the latest state-of-the-art YOLO model developed by Ultralytics, which can be used for object detection, image classification, and instance segmentation tasks [4]. The architecture of YOLOv8 builds upon the previous versions of YOLO algorithms and includes several key features and improvements.

The main components of YOLOv8 architecture are:

- Backbone: YOLOv8 utilizes a modified version of the CSPDarknet53 architecture, which consists of 53 convolutional layers and employs cross-stage partial connections.

- Head: The head part of the network is responsible for generating the final output and is responsible for predicting the bounding boxes, class labels, and abjectness scores.

- Neck: This part connects the backbone and the head, and in YOLOv8, it uses SPPF and New CSP-PAN structures.

- Adaptive Training: YOLOv8 uses adaptive training to optimize the learning rate and balance the loss function during training, leading to better model performance.

- Advanced Data Augmentation: YOLOv8 employs advanced data augmentation techniques such as MixUp and CutMix to improve the model's training performance.

- Customizable Architecture: YOLOv8's architecture is highly customizable, allowing users to easily modify the model's structure and parameters to suit their need.

Chapter 4

Implementation

4.1 Dataset Collection

The deep learning models were trained using our self-collected dataset and some data collected from the internet. We have collect data of table-top object for indoor scene classification. All the pictures of the dataset are captured manually using a RGB camera. Overall, the data set consists of 300 images for 100 image per class for 3 class i.e. ‘Cup’, ‘Can’, ‘Box’ mentation in fig-3.1, 3.2, and 3.3. The dataset is split into a training set (80%), a validation set (20%) for training and evaluating the model. Through the dataset acquisition process, we aimed to create a diverse and representative dataset that can accurately capture the characteristics of the target objects, and facilitate the development of an effective scene understanding model

Fig : 3.1 “BOX”

Fig: 3.2 “CAN”

Fig:3.3 “CUP”

4.2 Training Data

In the training phase of the project, we employed YOLOv8, a state-of-the-art object detection and segmentation model, to achieve multi-level segmentation and robust object detection. The training process involved preparing a labelled dataset with diverse examples of the target objects in various environmental conditions. I configured the YOLOv8 architecture, a deep neural network, by selecting appropriate hyperparameters and network configurations to suit the specific requirements of my application. To enhance the model's generalization capabilities, data augmentation techniques such as random scaling, cropping, and flipping were applied to the training dataset. During training, the model iteratively learned to predict bounding boxes and associated class probabilities for each object in the image. Training examples are shown in the fig : 4.1 , 4.2, 4.3 -

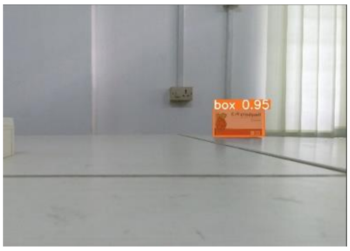


Fig:4.1 “Segmentate Box”

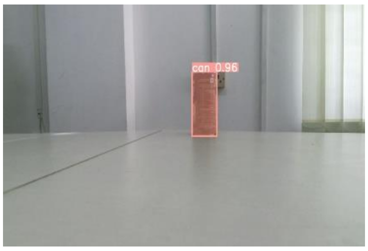


Fig:4.2 “Segmentate Can”

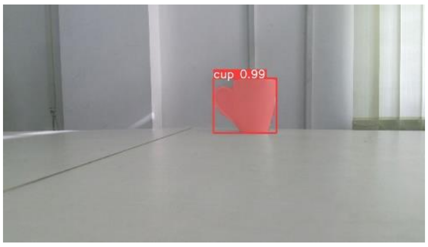


Fig:4.3 “Segmentate Cup”

Fig: 4.1.1

“Segmentate Box mask”

Fig: 4.1.2

“ Segmentate Can mask”

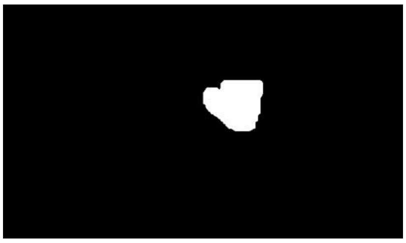
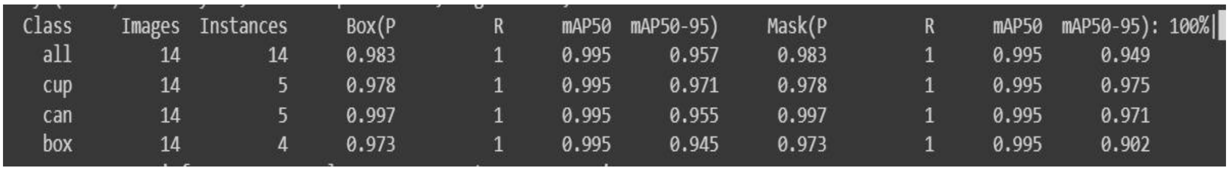


Fig: 4.1.3

“ Segmentate Cup mask”

Accuracy and prediction out put shown in fig : 4.1 below



```
 Class     Images  Instances      Box(P          R       mAP50  mAP50-95)     Mask(P          R       mAP50  mAP50-95): 100%|
   all         14         14      0.983          1       0.995      0.957      0.983          1       0.995      0.949
   cup         14          5      0.978          1       0.995      0.971      0.978          1       0.995      0.975
   can         14          5      0.997          1       0.995      0.955      0.997          1       0.995      0.971
   box         14          4      0.973          1       0.995      0.945      0.973          1       0.995      0.902
```

Fig: 4.2 “Output accuracy”

Chapter 5

Future Work

Till now we focused on multi-class object detection using YOLOv8, an important next step is to incorporate spatial context and relationship reasoning between objects through joint calibration with 2D lidar.

Currently, we rely only on visual data from the camera for object detection. By fusing it with depth information from lidar, we can encode important spatial cues to improve semantic segmentation and scene understanding.

The first task will be to properly calibrate the camera and 2D lidar so their data is aligned. We will explore techniques like extrinsic parameter calibration using a checkerboard and intrinsic calibration of lens distortion.

Once calibrated, the RGB image and depth map can be fused through methods like early and late fusion. The joint representation will be input to a semantic segmentation model like RefineNet. By incorporating lidar cues, model performance is expected to improve significantly.

Moving beyond segmentation, we also aim to add relationship reasoning, predicting how objects are positioned relative to each other like left, right, front, behind. This will provide a deeper understanding of indoor scene layouts.

Additional contextual modules can encode information like object co-occurrence relationships and likely positioning based on scene type. By jointly calibrating camera and lidar and encoding spatial context, we can achieve more meaningful semantic scene understanding.

In summary, we will extend our existing object detection to fused camera-lidar semantic segmentation and relationship reasoning between indoor objects and layouts. This will enable a more complete interpretation of complex indoor scenes

References

[1] X. H. J. L. Juan Li, “2D LiDAR and Camera Fusion in 3D Modeling of Indoor Environmen,” Springer, 2015.

[2] Z. Z. X. L. Z. H. Xu Song1, “Monocular camera and laser based semantic mapping system with temporal-spatial data association for indoor mobile robots,” Springer, 2023.

[3] S. B. T. S. B. E. Fanfei Chen, “Self-Learning Exploration and Mapping for Mobile Robots via Deep Reinforcement Learning.,” IEEE, 2019.

[4] W. W. Z. L. X. Z. D. Y. R. W. Xianyu Qi, “Object Semantic Grid Mapping with 2D LiDAR and RGB-D Camera for Domestic Robot Navigation.,” 2020.

[5] M. F. M. L. F. M. Harold F. Murcia(B), “3D Scene Reconstruction Based on a 2D Moving LiDAR.,” Springer, 2018.

[6] Z. M. Y. Z. Lu Jia, “A mobile Robot Mapping Method Integrating LiDAR and Depth Camera,” 2022.

[7] X. H. J. L. Juan Li, “2D LiDAR and Camera Fusion in 3D Modelling of Indoor Environment.,” Springer, 2023.

[8] H. W. ,. Z. W. ,. C. Z. ,. Y. Z. ,. T. T. Huazan Zhonga, “A survey of LiDAR and camera fusion enhancement.,” 2020.

[9] J. H. L. J. H. J. P. S. H. Y. S. K. G Ajay Kumar, “LiDAR and Camera Fusion Approach for Object Distance Estimation in Self-Driving Vehicles,” Springer, 2020.

[10] R. P. Qilong Zhang, “Extrinsic Calibration of a Camera and Laser Range Finder (improves camera calibration),” Springer, 2015.

[11] X. S. ·. Z. Z. ·. X. L. ·. Z. Huaidong, “Monocular camera and laser based semantic mapping,” Springer, 2023.

[12] P. R. Zhang Q, “Extrinsic calibration of a camera and laser range finder (improves camera,” IEEE, 2004.

[13] X. D. J. G. H. L. J. G. L. B. H. C. H Lou, “DC-YOLOv8,” *Small-Size Object Detection Algorithm Based on Camera Sensor ,* vol. 1, no. 21 May 2023, pp. 5-10, 2023.

[14] X. L. ·. Z. D. 1. ·. Y. Yang, “Recent progress in semantic image segmentation,” Springer, 2018.

[15] S. H. C. C. S. M. S. J. &. R. P. (. Lee, “ How deep learning extracts and learns leaf features for plant classification,” *Pattern Recognition,* pp. 71, 1-23, 2017.