KLE TECH KLE Technological University

Creating Value
Leveraging Knowledge

School of
Electronics and Communication Engineering

Mini Project Report

on

Sensor fusion for 3-D Object Detection

By:

1. **Puneet R K**	USN:01FE21BEC101
2. **Jayapriya A N**	USN:01FE21BEC134
3. **Pranav P**	USN:01FE21BEC096
4. **Shreya S Nadgir**	USN:01FE21BEC127

Semester: V, 2023-2024

Under the Guidance of

Prof Prabha C Nissimagoudar

K.L.E SOCIETY'S
KLE Technological University,
HUBBALLI-580031
2022-2023

SCHOOL OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that project entitled “ **Sensor fusion for 3-D Object Detection** ” is a bonafide work carried out by the student team of **”Puneet R K (USN:01FE21BEC101) Jayapriya A N (USN:01FE21BEC134) Pranav P (USN:01FE21BEC096)Shreya S Nadgir (USN:01FE21BEC127) ”**. The project report has been approved as it satisfies the requirements with respect to the mini project work prescribed by the university curriculum for BE (V Semester) in School of Electronics and Communication Engineering of KLE Technological University for the academic year 2023-2024.

Prabha C N	**Suneeta V Budihal**	**B. S. Anami**
Guide	**Head of School**	**Registrar**

External Viva:

Name of Examiners **Signature with date**

1.

2.

ACKNOWLEDGMENT

We would like to express our appreciation and gratitude to Dr.Suneeta V Budihal, Head of school, for providing us with an opportunity to learn and work on this project. We express our profound thanks to our mini project mentor prof. Prabha C N. We are deeply indebted to prof.Prabha C N for her vast array of technical knowledge and guidance regarding the technical aspects required to complete this project. Finally, we would like to thank our seniors, lab attenders and everyone else in the department from the ground up for their unwavering efforts, technical skills and support without which this project would not have been possible.

-The project team

ABSTRACT

In autonomous vehicles, the perception system is very crucial for understanding the surroundings and making safe decisions while driving, including identifying and interpreting objects, obstacles, and other vehicles.The perception system typically relies on a variety of sensors such as cameras, LiDAR (Light Detection and Ranging), radar, and ultrasonic sensors to gather data about the vehicle's surroundings. The accuracy and reliability of the perception system directly impact the safety and performance of autonomous vehicles. A robust perception system is essential for ensuring smooth and efficient navigation, and ultimately preventing accidents.

However, perception in autonomous vehicles is a challenging problem due to various factors such as adverse weather conditions, poor lighting, complex traffic scenarios, and certain limitations of the sensors. We fuse the data from multiple sensors to overcome the shortcomings of individual sensors. Sensor fusion refers to the integration and combination of information from multiple sensors to obtain a more accurate, reliable, and comprehensive understanding of the vehicle's surroundings.

The fusion of camera and LiDAR technology represents a significant advancement in perception systems, particularly in the realm of 3D object detection. This project is dedicated to the development of an algorithm that combines visual data captured by cameras with the precise depth information provided by LiDAR sensors, to obtain reliable information for 3d object detection. This fusion method is set to play a crucial role in enhancing the dependability of perception systems, in the field of self-driving cars, where precise and effective 3D object detection is crucial in varied and changing surroundings.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The use of LiDAR and camera data fusion for 3D object detection arises from the need to improve detection accuracy and adaptability in practical applications, as well as exploiting the complementary capabilities present in each sensor modality [6].When it comes to gathering spatial data, LiDAR systems are precise, producing dense point clouds that accurately depict the geometry, structure, and distance of their immediate surroundings providing the depth measure of the surroundings. However, unfavorable weather conditions like intense rain or fog hinder the functionality of LiDAR sensors and reduce their efficacy. On the other hand, precise object detection is possible by RGB cameras, which offer rich color and texture information. While cameras are good at collecting visual details, they often have trouble resolving occlusions and precisely measuring depth in complicated situations[3]. Through the fusion of data collected from LiDAR and the camera, the combined system makes use of the special benefits of both modalities and depends less on a single sensor for perception, to overcome the limitations of each sensor and achieve better detection performance. Accurate object detection in three dimensions is made possible by the combination of LiDAR and camera data, which provides a more thorough picture of the scene. This will ultimately improve the deployment of autonomous systems in real-world scenarios and expand their capabilities.

1.2 Objectives

- Integrate the colour and texture information from the camera with the spatial data from LiDAR point cloud to obtain more reliable information.

- To use complementary capabilities of each of the sensors to overcome the limitations of the individual sensors.

- Developing algorithms that can attain real-time performance, crucial for applications like driverless cars, where quick decision-making is essential.

- Perform 3D object detection to attain improved safety features and facilitate the development of more advanced autonomous driving technologies, ultimately contributing to safer driving experiences.

1.3 Literature survey

1. **3D vehicle detection based on LiDAR and camera fusion** (Cai, Yingfeng, et al., 2019)

To improve detection accuracy, a unique approach to 3D vehicle identification in their work that integrates data from cameras and LiDAR. They provide a network architecture for object recognition that efficiently combines data from RGB pictures and LiDAR point clouds. Using a combination of RGB image fusion and LiDAR point cloud projection, the technique creates high-resolution feature maps that are ready for further processing. The method produces dependable 3D vehicle proposals and conducts oriented 3D box regression for precise extent and orientation prediction by utilizing a 3D proposal network and region proposal fusion. When compared to current techniques, the experimental findings on the KITTI dataset show improved performance for 3D detection accuracy [1].

2. **Real-time object detection using LiDAR and camera fusion for autonomous driving.** (Liu, H., Wu, C., & Wang, H., 2023)

The proposed object detection algorithm involves preprocessing raw point clouds from the KITTI dataset by converting them into depth images, facilitating easier processing and real-time performance enhancement. This conversion entails multiple steps, including projection, transfer, and transformation, to align the data from LiDAR to the pixel coordinate system. The algorithm employs a Siamese network architecture with parallel branches to process both depth and RGB images, extracting feature maps for subsequent analysis. Feature-layer fusion is introduced to fuse the extracted feature maps from multi-modality data, enhancing object detection accuracy. While demonstrating the effectiveness of the conversion process by overlaying depth images onto RGB images, the algorithm faces challenges due to the large data volume resulting from feature fusion, which could strain the GPU. Nonetheless, the proposed approach aims to optimize object detection efficiency for autonomous vehicles by streamlining data processing and implementing effective fusion strategies[5].

3. **Camera-LiDAR Multi-Level Sensor Fusion for Target Detection at the Network Edge** (Mendez, J.; Molina, M.; Rodriguez, N.; Cuellar, M.P.; Morales, D.P., 2021)

Detecting targets by combining LiDAR and camera data, the proposed Multi-Level Sensor Fusion (MLSF) Network makes use of the SSD structure to provide high-performance accuracy while using less memory. To overcome issues relating to the unordered nature of 3D point clouds, LiDAR depth maps are created from raw data to depict surfaces in 2D pictures. This preprocessing keeps relevant characteristics while drastically lowering memory use. The MLSF model combines feature maps at different levels by using fusion layers and distinct CNNs for camera pictures and LiDAR depth maps. Multi-level fusion is facilitated by fusion layers that concatenate feature maps from both data sources and apply convolutional filters to build deeper feature maps. By reducing the number of layers and quantizing parameters to 8-bit integers, the network is optimised for edge devices. The complete pipeline consists of quantizing input data, creating LiDAR depth maps, aligning data from LiDAR and camera sensors, and using the MLSF model to detect targets. These methods satisfy the requirements of edge devices like as the Google Coral TPU Development Board, and at the same time provide efficient target identification[7].

1.4 Problem statement

Develop a robust and accurate 3D object detection ”system” for autonomous vehicles by effectively combining the complementary information from camera and LiDAR sensors.

1.5 Application in Societal Context

The integration of LiDAR and camera sensor fusion for 3D object detection presents transformative applications across various societal domains. In autonomous vehicles, this technology enhances safety by accurately identifying objects like pedestrians, vehicles, and road signs, empowering vehicles to make informed decisions in real-time and navigate complex environments with greater precision.[**?**] Moreover, in urban planning, LiDAR and camera fusion facilitates detailed 3D mapping, aiding in optimizing transportation networks, improving city layouts, and enhancing the overall livability and sustainability of urban areas. Furthermore, in surveillance and security systems, the technology enables precise detection and tracking of objects in three-dimensional space, bolstering security monitoring efforts and ensuring the safety of public spaces, critical infrastructure, and sensitive facilities. Additionally, LiDAR and camera fusion finds application in environmental monitoring and conservation by providing detailed insights into natural landscapes, ecosystems, and wildlife habitats, facilitating biodiversity monitoring, ecological assessment, and conservation planning. In summary, the utilization of 3D object detection using LiDAR and camera sensor fusion drives advancements that enhance safety, efficiency, and sustainability across diverse societal contexts.

1.6 Organization of the report

- **Chapter 1: Introduction**
- **Chapter 2: System Design**
- **Chapter 3: Algorithm and Flowchart Insight**
- **Chapter 4: Results and Discussion**
- **Chapter 5: Conclusion and Future Scope**

Chapter 2

System design

In our 3D object detection system design, we leverage data collected from a comprehensive sensor suite provided by the nuScenes dataset. This dataset comprises 1 LiDAR sensor, 5 RADAR sensors, and 6 monocular cameras. The LiDAR sensor, specifically the Velodyne HDL-64E LiDAR, is positioned on the vehicle's roof, offering a panoramic 360-degree horizontal field of view and a 26.8-degree vertical field of view. It enables precise 3D point cloud generation with a range spanning from 0.9m up to 12m. Additionally, the system incorporates monocular cameras, each boasting a 120-degree horizontal field of view and a 35-degree vertical field of view, enhancing the visual perception capabilities of the system. It's important to note that we are utilizing the nuScenes mini dataset, a subset of the larger nuScenes dataset, for our system development. Despite being a reduced version, the nuScenes mini dataset still provides rich and diverse sensor data, including LiDAR point clouds and camera images, along with accurate ground truth annotations for object detection tasks. By leveraging the nuScenes mini dataset, we can efficiently develop and evaluate our 3D object detection system while benefiting from the comprehensive sensor suite and high-quality annotations provided by the dataset.

2.1 Functional block diagram

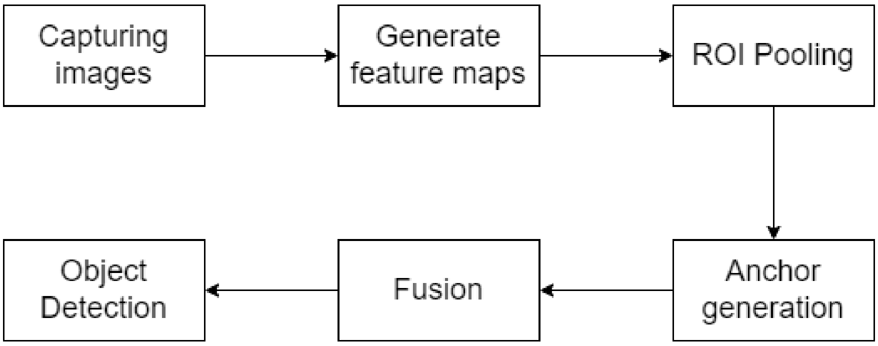

Figure 2.1: Functional block diagram for Camera and LiDAR sensor fusion

1. Capture Images: The process starts by capturing images of a city street using a camera. These images are likely to be RGB images, which means they capture the red, green, and blue color channels of light.

2. Generate Bird's-Eye View (BEV) Map: The next step is to generate a bird's-eye view (BEV) map. A BEV map is a type of image that shows the world from above, as if it were being viewed from a bird's eye.

3. Generate Feature Maps: The next step is to generate feature maps from the RGB images and BEV map. Feature maps are a type of image representation that captures the most important features of an image, such as edges, shapes, and textures. There are many ways to generate feature maps, but a common approach is using a convolutional neural network (CNN).

4. Region of Interest (ROI) Pooling: Once the feature maps have been generated, the next step is to perform ROI pooling. ROI pooling is a technique that is used to select specific regions of interest from the feature maps. These regions of interest are likely to contain the most important information for object detection.

5. Anchor Generation: The next step is to generate anchors. Anchors are small boxes that are placed at different locations on the BEV map. These anchors will be used to predict the presence and location of objects in the image.

6. Fusion: The next step is to fuse the information from the BEV map and the anchors. This fusion can be done in several ways. A technique that is used to remove redundant bounding boxes and keep only the most confident predictions.

7. Object Detection: The final step is to perform object detection. This involves using the fused information from the BEV map and the anchors to predict the presence and location of objects in the image. The output of this step is a list of bounding boxes, each of which corresponds to a detected object.

2.2 Final design

In creating a design that fits our project needs within our time and budget limits, we carefully look at different options. We consider what resources we have and what restrictions we need to work with. After comparing our choices, we pick the best design based on how well it works and how easy it is to put into action. We think about things like how the design fits our needs and how manageable it is to make happen. Choosing the right design helps us run our project smoothly and efficiently, making the most of what we have while also keeping things simple and practical.

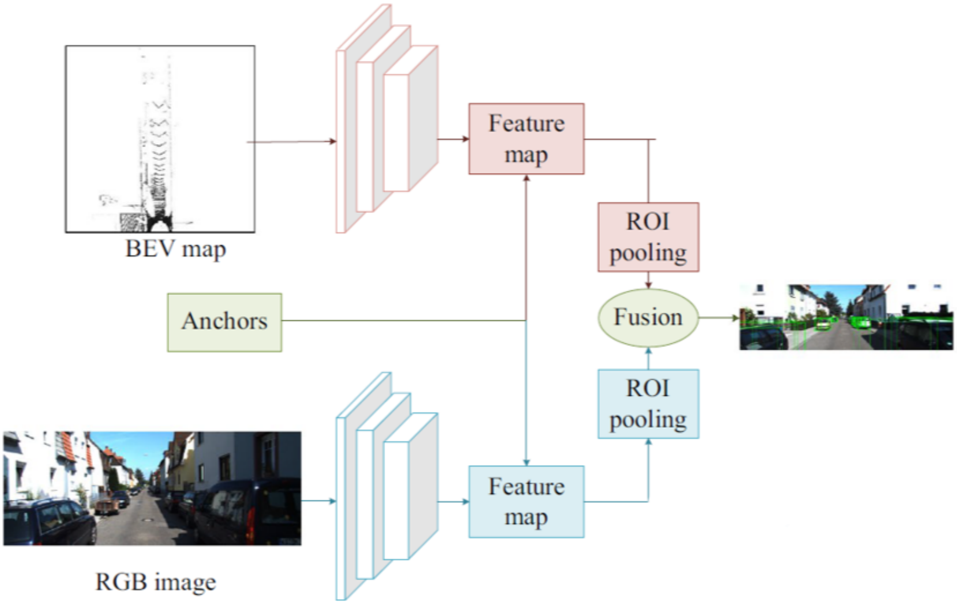

Figure 2.2: Final design for Camera and LiDAR sensor fusion

Chapter 3

Implementation details

3.1 Specifications and final system architecture

3.2 Algorithm

3.2.1 Input

- RGB image
- LiDAR point cloud

3.2.2 Preprocess LiDAR

Generate BEV map

- Project each LiDAR point onto the ground plane based on its height.
- Accumulate the points in each grid cell of the BEV map.
- For each cell, calculate:
  - Maximum height of points in the cell.
  - Number of points in the cell.

3.2.3 Feature Extraction

- Extract features from the BEV map and RGB image using the VGG-16 architecture.
- Apply a 1x1 convolution to each feature map to fuse low-level and high-level features.
- Upsample the lower-resolution feature maps to match the size of the higher-resolution ones. Perform element-wise addition to merge the upsampled and original feature maps.
- Apply a 3x3 convolution to each merged feature map to generate the final feature maps.

3.2.4 3D Proposal Network

- Generate a set of 3D anchor boxes with different sizes and orientations at predefined locations in the BEV map and RGB image.
- For each anchor box:
  - Project the anchor box onto the BEV and RGB feature maps.
  - Perform ROI pooling on the feature maps to extract a fixed-size feature vector for the anchor.
  - Fuse the ROI pooled features from BEV and RGB using element-wise mean.
  - Classify the anchor as object or background based on its Intersection Over Union (IOU) with ground-truth boxes.
  - If the anchor is classified as an object:
    - * Perform 3D box regression to refine the position and dimensions of the anchor box.

$$\mathbf{t}_i = (t_x, t_y, t_z, t_l, t_w, t_h) \tag{3.1}$$

Equation 3.1 is the offset of the predict box relative to the 3D anchor box

$$\mathbf{t}_i^* = (t_x^*, t_y^*, t_z^*, t_l^*, t_w^*, t_h^*) \tag{3.2}$$

Equation 3.2 is the off set of the ground-truth box relative to the 3D anchor box

$$t_x = \frac{x_g - x_a}{d_a} \tag{3.3}$$

$$t_y = \frac{y_g - y_a}{d_a} \tag{3.4}$$

$$t_z = \frac{z_g - z_a}{h_a} \tag{3.5}$$

$$t_l = \log\left(\frac{l_g}{l_a}\right) \tag{3.6}$$

$$t_w = \log\left(\frac{w_g}{w_a}\right) \tag{3.7}$$

$$t_h = \log\left(\frac{h_g}{h_a}\right) \tag{3.8}$$

$$t_x^* = \frac{x_p - x_a}{d_a} \tag{3.9}$$

$$t_y^* = \frac{y_p - y_a}{d_a} \tag{3.10}$$

$$t_z^* = \frac{z_p - z_a}{h_a} \tag{3.11}$$

$$t_l^* = \log\left(\frac{l_p}{l_a}\right) \tag{3.12}$$

$$t_w^* = \log\left(\frac{w_p}{w_a}\right) \tag{3.13}$$

$$t_h^* = \log\left(\frac{h_p}{h_a}\right) \tag{3.14}$$

$$d_a = \sqrt{l_a^2 + w_a^2} \tag{3.15}$$

where

$$(x_g, y_g, z_g, l_g, w_g, h_g) \tag{3.16}$$

is the ground-truth box,

$$(x_a, y_a, z_a, l_a, w_a, h_a) \tag{3.17}$$

is the 3D anchor box

$$(x_p, y_p, z_p, l_p, w_p, h_p) \tag{3.18}$$

is the predict box

$$\mathbf{d}_a \tag{3.19}$$

is the diagonal length of the 3D anchor box.

- Apply Non-Maximum Suppression (NMS) to select the top proposals based on their classification scores.

3.2.5 Region Proposal Fusion

- For each proposal:
  - Extract its ROI pooled features from the BEV and RGB feature maps.
  - Fuse the features using element-wise mean.

3.2.6 3D Box Regression

- Use the fused features from each proposal to predict:
  - Classification score (vehicle or background).
  - Offset values for the box center, length, width, and two heights relative to the anchor box.
- Decode the predicted offsets to obtain the final 3D bounding boxes.

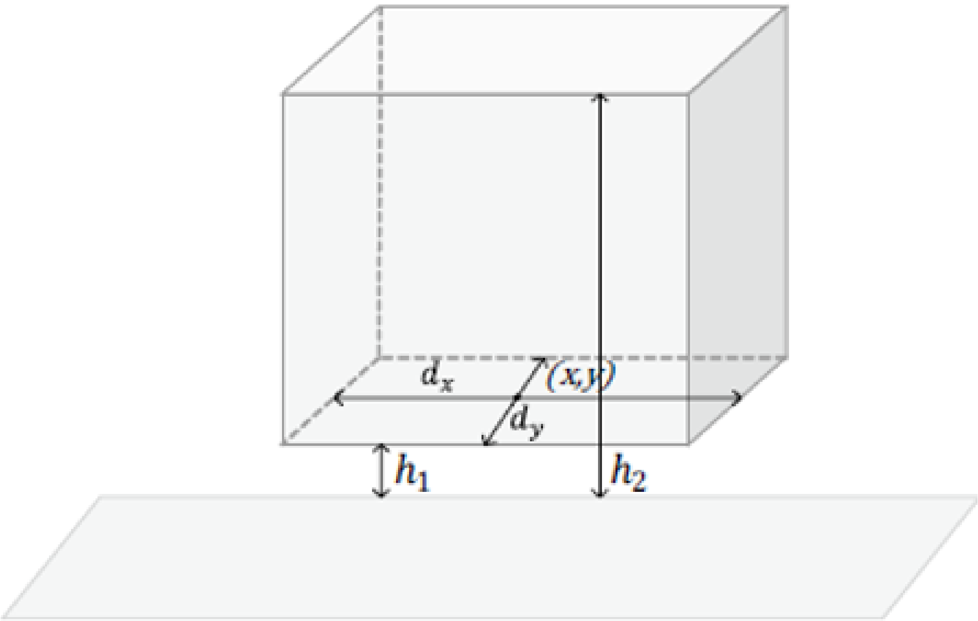

Figure 3.1: Encoded 3D bounding box
The regression targets are encoded by $\Delta x, \Delta y, \Delta d_x, \Delta d_y, \Delta h_1, \Delta h_2$

3.2.7 Output

The list of predicted 3D bounding boxes with their corresponding classification scores.

3.3 Flowchart

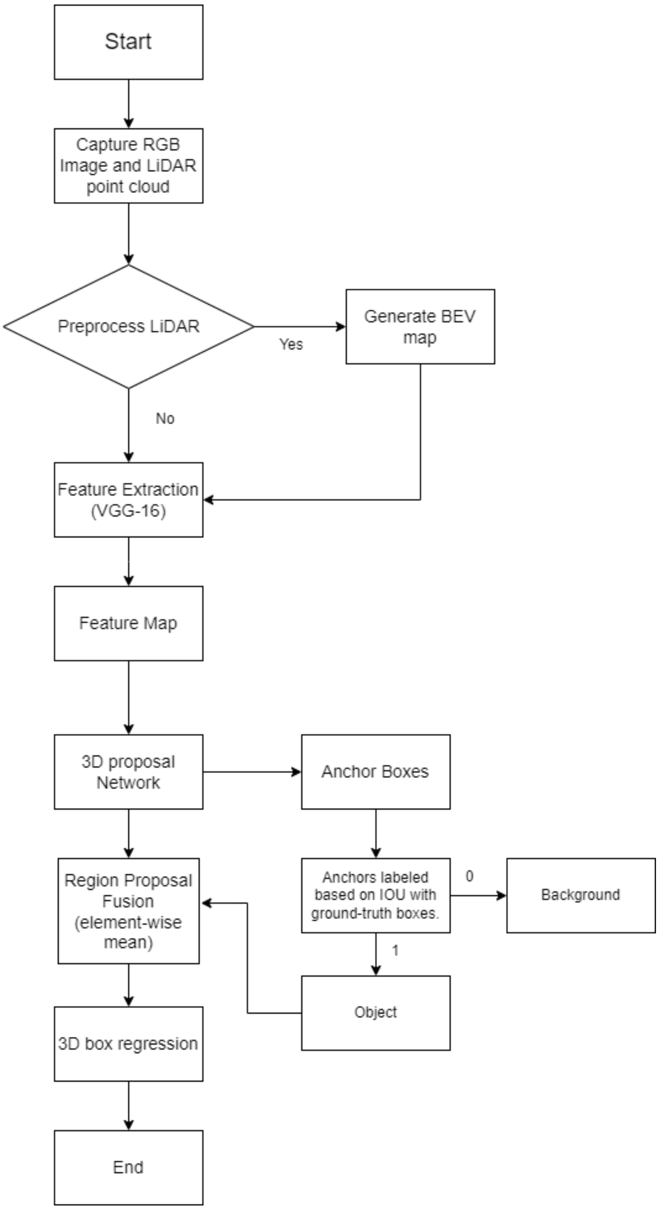
Start
Capture RGB Image and LiDAR point cloud
Preprocess LiDAR
Yes
Generate BEV map
No
Feature Extraction (VGG-16)
Feature Map
3D proposal Network
Anchor Boxes
Region Proposal Fusion (element-wise mean)
Anchors labeled based on IOU with ground-truth boxes.
0
Background
1
Object
3D box regression
End

Figure 3.2: Flowchart for algorithm

Chapter 4

Results and discussions

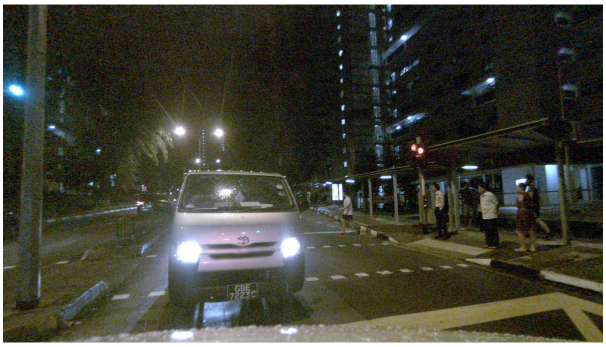

Figure 4.1: Input RGB image

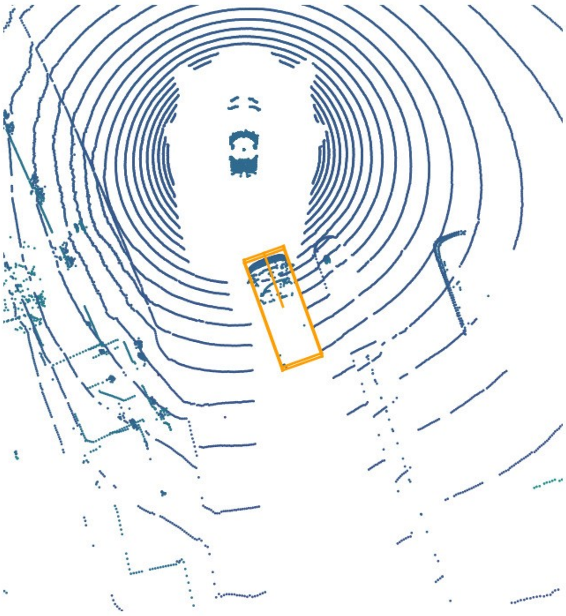

Figure 4.2: LiDAR point cloud

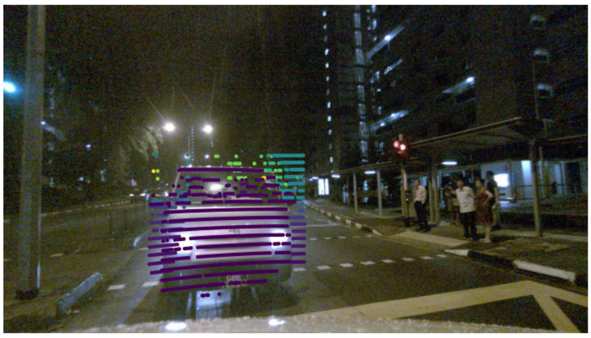

Figure 4.3: Region of interest for fusion

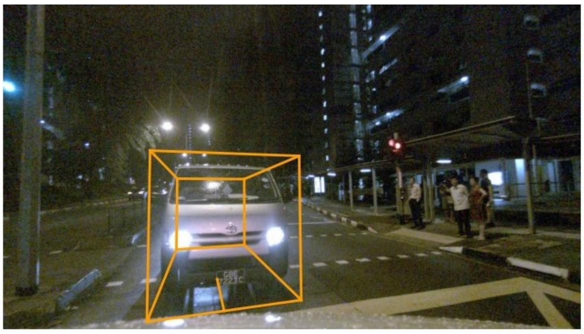

Figure 4.4: 3D box for the detected object

4.1 Result Analysis

Table 4.1: Performance comparison of different methods on the NuScenes dataset.

Method	Average Precision 3D (%)			Average Precision BEV (%)		
	Easy	Moderate	Hard	Easy	Moderate	Hard
RT3D [8]	21.27	23.49	19.81	42.10	54.68	54.68
Stereo R-CNN[4]	34.05	49.23	28.39	43.89	61.67	36.44
A3DODWTD [2]	56.81	59.35	50.51	72.86	76.65	76.65
This project	**68.59**	**63.72**	**53.34**	**73.56**	**66.75**	**58.78**

The proposed method leverages a fusion of BEV maps and RGB images with 3D anchor boxes and region proposal fusion. It outperforms other techniques in 3D vehicle detection on the NuScenes dataset. For Easy and Moderate difficulties, our method achieves an Average Precision (AP) of 68.59 % and 63.72 %, respectively, surpassing previous methods of 56.81 % and 59.35 %. Our method achieves an AP of 63.91 %, better than others. These findings show how effective our approach is for 3D vehicle detection, especially in challenging situations with different object positions and blockages.

Chapter 5

Conclusions and future scope

5.1 Conclusion

In conclusion, our study presents a robust framework for 3D object detection through the fusion of camera and LiDAR sensors. By leveraging the complementary strengths of both modalities, we achieve significant results in detecting objects reliably. Our approach addresses various challenges inherent in 3D object detection, including occlusions and various environmental conditions. Through rigorous experimentation and evaluation, we have demonstrated the superior performance of our method compared to existing techniques, achieving higher accuracy and precision across different difficulty levels.

Furthermore, our framework not only enhances the perception capabilities of autonomous systems but also contributes to improving safety and efficiency of autonomous driving. The integration of camera and LiDAR data enables our system to capture detailed spatial information while maintaining computational efficiency, making it suitable for real-time applications.

5.2 Future scope

Looking ahead, future research directions may include exploring advanced fusion algorithms, incorporating semantic information for object recognition, and enhancing the robustness of the system against challenging scenarios. Additionally, efforts towards standardization and benchmarking will be crucial for facilitating comparisons and advancements in the field of 3D object detection using camera and LiDAR fusion. Overall, our work lays a solid foundation for further innovation and development in this rapidly evolving area of computer vision and autonomous systems.

Bibliography

[1] Yingfeng Cai et al. 3d vehicle detection based on lidar and camera fusion. *Automotive Innovation*, 2:276–283, 2019.

[2] Zeming Cai, Qi Fan, Rogerio S Feris, et al. A unified multi-scale deep convolutional neural network for fast object detection. *Comput. Vis.*, 9908:354–370, 2016.

[3] L. Guan, Y. Chen, G. Wang, and X. Lei. Real-time vehicle detection framework based on the fusion of lidar and camera. *Electronics*, 9(451), 2020.

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. Feature pyramid networks for object detection. *CVPR*, 11(2):936–944, 2016.

[5] H. Liu, C. Wu, and H. Wang. Real time object detection using lidar and camera fusion for autonomous driving. *Sci Rep*, 13:8056, 2023.

[6] J. Madake, R. Rane, R. Rathod, A. Sayyed, S. Bhatlawande, and S. Shilaskar. Visualization of 3d point clouds for vehicle detection based on lidar and camera fusion. In *2022 OITS International Conference on Information Technology (OCIT)*, pages 594–598. IEEE, 2022.

[7] J. Mendez, M. Molina, N. Rodriguez, M.P. Cuellar, and D.P. Morales. Camera-lidar multi-level sensor fusion for target detection at the network edge. *Sensors*, 21(3992), 2021.

[8] G Yebo, Y Minglei, S Zhenguo, et al. The applications of decision-level data fusion techniques in the field of multiuser detection for ds-uwb systems. *Sensors*, 15(10):24771–24790, 2015.