LiDAR-Camera Fusion 3D Object Detection Using Deep Learning Algorithms

[1]Dr. Sachin Singh, [2]Arjun Singh. [3]Dr.Khoob Singh, [4]Vibhuti Sharma, [5]Mr Subodh Rastogi, [6]Piyush Rastogi.

[1,4,5,6] MIT Moradabad, India, [2]GNIOT Greater Noida, India. [3]Maya Devi University Dehradun, India

[1,2] Dept. of CSE, [4,5] Dept. of Computer Applications, [3] Dept. School of Engineering,[6] Dept of CSE- AIML

ABSTRACT

The rapid progress in autonomous driving technology has highlighted the need for accurate and reliable systems to detect objects, especially in 3D environments. This study aims to improve 3D object detection by combining data from LiDAR and cameras. LiDAR sensors are known for their ability to measure distances precisely, while cameras provide detailed visual information. By merging these two data sources, the goal is to create a system that can detect objects more accurately and reliably, even in complex and dynamic driving conditions. The research involves three main objectives: developing a method for combining sensor data, using advanced deep learning techniques to analyse this combined data, and validating the system through experiments. To achieve this, the project uses a pre-trained YOLOv5 model to identify objects in 2D camera images. The identified objects are then mapped into 3D space using LiDAR data, which is carefully aligned and calibrated with the camera images. The fusion process links the depth information from the LiDAR point cloud with the visual data from the camera, enabling precise 3D positioning of objects. The results show that combining LiDAR and camera data significantly enhances the accuracy of 3D object detection. Tests comparing estimated object distances with actual measurements reveal only minor differences, confirming the system's effectiveness and reliability. This work demonstrates the importance of integrating multiple sensors to improve the performance of perception systems in autonomous vehicles. This study contributes to the field of autonomous driving by presenting a validated system for LiDAR-Camera fusion. The findings emphasise how sensor fusion can enhance the robustness and precision of object detection systems. Future research could explore how to optimise the system for challenging weather conditions, incorporate additional sensors like RADAR, and leverage more advanced deep learning models to further advance autonomous driving technology.

Keywords: Autonomous Driving, 3D Object Detection, Distance Measurement, Sensor Fusion, LiDAR-Camera Integration, Deep Learning, CNNs

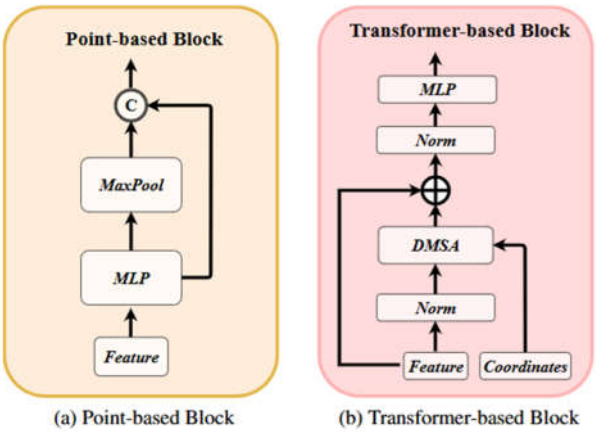
Introduction

Autonomous driving technology is poised to revolutionise transportation, making travel safer, more efficient, and more convenient. A crucial aspect of autonomous vehicles is their ability to accurately perceive and interpret their surroundings to navigate safely with little or no human intervention. This thesis focuses on developing a system for 3D object detection by combining data from LiDAR (Light Detection and Ranging) and cameras. By integrating these two sensing technologies, the aim is to enhance the reliability and precision of object detection, even in complex and dynamic driving environments. This chapter outlines the problem addressed in the thesis, offering readers a clear understanding of its purpose. In recent years, deep learning-based 2D object detection has garnered significant attention. Many researchers have extended these approaches to 3D object detection using LiDAR point clouds. However, LiDAR-generated point clouds are sparse and irregular, which creates challenges for accurate detection. To overcome this, some studies transform point clouds into 2D representations, such as front-view images, bird's-eye view (BEV) images, or structured voxel grids. These transformations allow 2D convolutional neural networks to extract features, but they often result in some loss of 3D information, particularly in distant regions. Point-based methods that directly process LiDAR point clouds offer another approach, often using multilayer perceptron's to extract features. While these methods can provide more detailed information, they are computationally intensive. BEV-based methods are faster but still lose some information during the conversion process. This thesis mitigates information loss by incorporating RGB-D images, which combine 2D RGB image data with depth information. RGB images are rich in texture and high-resolution details, making them well-suited for detecting small objects. However, when using only monocular or even stereo images, accurately estimating depth remains challenging. To address these

limitations, some studies have explored combining 2D image data with LiDAR point clouds. Traditional fusion techniques, such as simple concatenation or elementwise averaging of features, often result in suboptimal 3D detection accuracy. This thesis adopts an advanced Region of Interest (ROI) attention fusion mechanism, which enables more effective integration of features from different sensor modalities. Previous research has also explored two-stage and one-stage frameworks for 3D object detection. Two-stage methods, like MV3D and AVOD, generate initial 3D proposals in the first stage and refine them in the second, achieving high accuracy but at the cost of greater computational time. One-stage methods are faster but often sacrifice accuracy because they lack a refinement stage. This study improves the performance of one-stage models by incorporating a global feature attention (GFA) mechanism, which enhances the representation of global features, thereby boosting detection accuracy. By addressing these challenges, this thesis aims to contribute to the development of more robust and accurate 3D object detection systems, advancing the capabilities of autonomous vehicles in real-world scenarios.

Method

This chapter describes the research methodology used in this thesis, explaining the reasoning behind the chosen approach and how it aligns with the research objectives. It outlines why the selected methodology is well-suited to address the research question and discusses the rationale for choosing it over other alternatives, ensuring that the decision is grounded in scientific principles and tailored to the specific requirements of the problem. Additionally, the chapter examines how the chosen methodology impacts the validity and reliability of the research findings. The research adopts an implementation-based methodology, which involves designing, developing, and testing new solutions, such as algorithms or techniques, to evaluate their performance against existing methods. As outlined by Berndtsson et al. (2007), this approach focuses on demonstrating the practical advantages of a proposed solution through implementation and comparison. The aim is to showcase measurable improvements and validate the solution's effectiveness in real-world applications. In this study, the proposed solution entails developing a system that combines data from LiDAR and cameras to detect objects and estimate their distances. This system is evaluated using real-world data to ensure accuracy, reliability, and performance. The evaluation includes comparisons with existing methods, as discussed in Section 2.3.1, to establish the system's advantages and high confidence in its results. The overall workflow of the RCBEVDet system is illustrated in Figure 2. The process begins with multi-view images, which are passed through an image encoder to extract features. A view transformation module then converts these image features into bird's-eye view (BEV) features. Simultaneously, radar point cloud data, aligned with the image data, is processed by the RadarBEVNet to generate radar BEV features. These two sets of features—image BEV and radar BEV—are then combined using a Cross-Attention Multi-Layer Fusion module. The final fused BEV features are utilised for the 3D object detection task, ensuring accurate and robust results. This methodology ensures that the system is rigorously tested and validated, highlighting its potential to advance object detection in autonomous driving applications.



(a) Point-based Block (b) Transformer-based Block

Future Scope

The future potential of LiDAR-Camera fusion for 3D object detection is immense, especially as advancements in autonomous systems, robotics, and smart transportation continue to demand highly accurate and dependable perception technologies. Combining data from LiDAR and cameras leverages their respective strengths—cameras provide high-resolution imagery and colour details, while LiDAR delivers precise depth measurements. Together, they enable more reliable and context-aware object detection. Future research can focus on optimising fusion techniques using advanced deep learning frameworks to improve accuracy and efficiency. Efforts can also aim at enhancing real-time processing to enable deployment on edge devices and making the system more resilient in challenging conditions, such as fog, rain, or low light. Beyond autonomous vehicles, this technology has applications in emerging areas like mixed reality and digital twins, offering improved spatial understanding for urban planning, construction, and the entertainment industry. Another promising avenue is reducing the cost and energy consumption of hardware without compromising performance, making these systems more accessible for broader adoption in consumer and industrial settings. Integration with advanced driver assistance systems (ADAS) and smart city infrastructure can significantly enhance road safety, traffic management, and autonomous mobility, further driving the adoption of this technology in shaping the future of intelligent transportation.

2. Related Work

2.1. Camera-based 3D Object Detection

Detecting objects in 3D space using only camera images presents significant challenges due to the limited depth information compared to LiDAR or radar systems. However, researchers have made substantial progress in overcoming this limitation through various approaches. These include estimating depth from images, leveraging geometric constraints and shape priors, designing specialised loss functions, and optimising detection and reconstruction jointly. The availability of multi-view camera datasets has further advanced 3D object detection by enabling the development of Multiview-based methods. These methods fall into two main categories: geometry-based and transformer-based approaches. Geometry-based methods, such as the Lift-Splat-Shoot (LSS) model, transform image features from multiple viewpoints into 3D voxel or bird's-eye view (BEV) representations. LSS uses a depth estimation network to compute depth distributions and a context vector for each image, combining them to generate 3D features along the camera's perspective rays. BEVDet improves upon LSS by directly detecting 3D objects within the BEV feature space, while BEVDepth enhances depth estimation accuracy with explicit depth supervision. Building on these, BEVDet4D aligns BEV features from past image frames to improve velocity predictions and overall detection performance. The RCBEVDet system, highlighted in this study, builds on these advancements. The pipeline begins by encoding features from multi-view images and transforming them into BEV representations, generating the image BEV features. Simultaneously, radar point clouds are processed through RadarBEVNet to extract radar BEV features. These BEV features from cameras and radar are dynamically aligned and combined using a cross-attention multi-layer fusion (CAMF) module. The resulting semantically rich, fused BEV features are then utilised for accurate 3D object detection tasks, offering improved precision and reliability.

2.2. Radar-camera 3D Object Detection

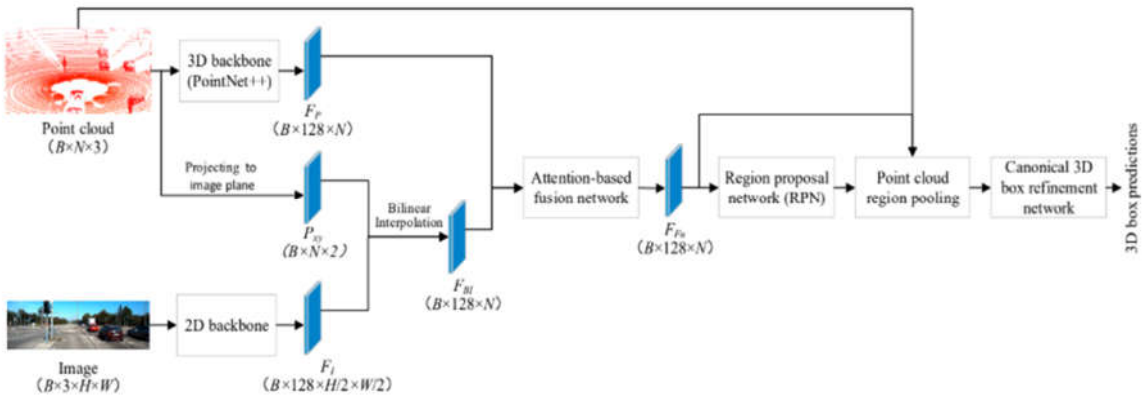
Millimetre-wave radar is a commonly used sensor in autonomous vehicles for 3D object detection due to its cost-effectiveness, long-range capabilities, and the ability to provide Doppler velocity measurements that are not impacted by adverse weather conditions. However, the sparse nature of radar data and its lack of semantic information make radar-only 3D object detection challenging. As a result, radar is typically used in combination with other sensors, such as cameras, in multi-modal 3D object detection systems. Recently, the combination of millimetre-wave radar with multi-view cameras has garnered significant attention, as these sensors complement each other well—radar offers depth and velocity information, while cameras provide rich visual data. Various methods have been developed to combine radar and camera data to enhance 3D object detection. For example, Radar Net employs a multi-level fusion approach that improves the detection of distant objects and reduces velocity measurement errors. CenterFusion generates initial 3D detections from camera images and refines them by associating radar features. CRAFT introduces a proposal-level fusion approach that uses a Soft-Polar Association and Spatio-Contextual Fusion Transformer to

efficiently exchange information between radar and cameras. RADIANT estimates the offset between radar echoes and object centres, using radar depth information to enhance camera features. CRN creates radar-augmented images with depth information from radar, which are then processed through multi-view transformation using a cross-attention mechanism to address misalignment and information gaps between radar and camera data. RCFusion employs a radar pillar net to generate radar pseudo-images and combines radar and camera bird's-eye-view features with a weighted fusion module. In contrast to these approaches, RCBEVDet introduces a specialised RadarBEVNet for efficient radar BEV feature extraction and a Cross-Attention Multi-Layer Fusion module, which ensures robust alignment and fusion of features from both radar and camera data, resulting in improved integration and more accurate and reliable 3D object detection in autonomous vehicles.

Implementation

This chapter describes the approach used in this thesis to tackle the key challenges in 3D object detection for autonomous vehicles by combining LiDAR and camera data. It explains the step-by-step process involved in gathering data, developing the detection model, merging sensor data, and associating depth information to ensure accurate 3D object detection.

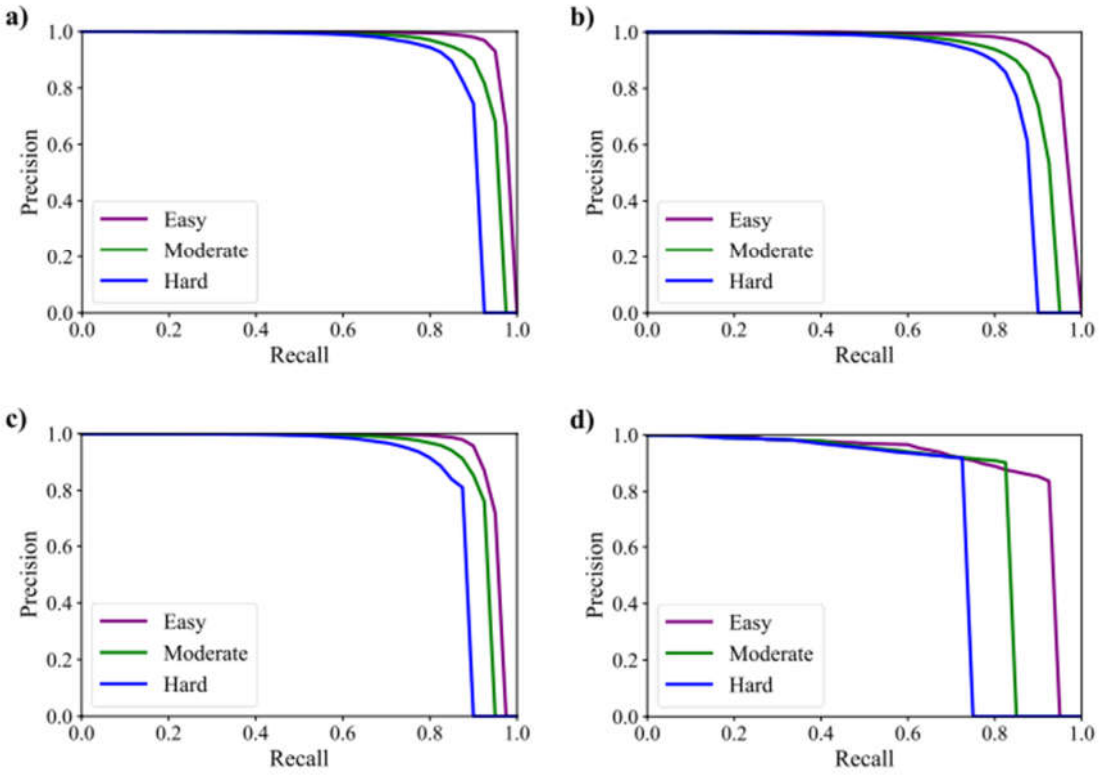
FuDNN for 3D Object Detection



A deep learning model called FuDNN, based on point CNN, is designed for 3D object detection. Its architecture includes several components: a 2D backbone, a 3D backbone, an attention-based fusion sub-network, a region proposal network (RPN), and a 3D box refinement network. The 2D backbone extracts 2D features from camera images, while the attention-based fusion sub-network combines these 2D features with 3D features from LiDAR data, extracted by Point Net++. The RPN generates 3D object proposals, and the 3D box refinement network fine-tunes the 3D object locations. The inputs to FuDNN are point clouds and RGB images. The images are processed as a matrix of dimensions B × 3 × H × W, where B is the batch size, H is the height, and W is the width of the image. The point clouds are represented as a matrix of B× 3× N, where N is the number of LiDAR points. The 2D backbone starts with a convolutional layer (Conv1) with 128 kernels of size 7 × 7, a stride of 1, which outputs a matrix of size B × 128 × H × W. Following this, batch normalisation (BN1) is applied to speed up network training and convergence, as demonstrated by Ioffe and Szegedy. A ReLU activation (ReLU1) is used to avoid the vanishing gradient problem, and then a max-pooling layer (S1) with a 2 × 2 kernel reduces the size of the feature map to B × 128 × H/2 × W/2. This structure is repeated twice: a second convolutional layer (Conv2) with 256 kernels of size 5 × 5 is followed by batch normalisation (BN2) and ReLU activation (ReLU2). The third convolutional layer (Conv3) has 128 kernels of size 3 × 3, followed by ReLU3. The output of the 2D backbone is the image feature matrix, denoted as FI, with the shape B × 128 × H/2 × W/2.

COMPARISON WITH TABLE

Method	Pub. Year	Stage(s)	Number of Parameter	Runtime (Ms)	3D (%)				BEV (%)			
					E	M	H	mAP	E	M	H	mAP
MV3D [4]	2017		-	360	71.29	62.678	56.56	63.51	86.55	78.10	76.67	
F-Point Net [26]	2017		-	170	83.76	70.92	63.65	72.78	88.16	84.02	76.44	
PC-CNN [27]	2018		-	500	57.63	51.74	51.39	53.59	83.61	77.36	69.61	
AVOD [14]	2018	Two	38,073,528	80	83.11	74.02	67.84	74.99	-	-	-	
AVOD-FPN [14]	2018		-	100	84.41	74.44	68.65	75.83	89.37	86.09	79.13	
MVX-Net [28]	2019		-	150	85.50	73.30	67.40	75.40	89.50	86.90	79.00	
MCF3D [21]	2019	Three	-	160	84.11	75.19	77.23	77.84	88.82	86.11	79.31	
AVOD-SSD [15]	2018		13,399,918	90	82.36	72.92	74.12	74.12	89.00	85.08	78.31	
Cont-Fuse [29]	2019	One	-	60	**86.32**	73.25	75.79	75.79	**95.44**	**87.34**	82.43	
Complex-Retina [16]	2019		-	90	78.62	72.77	72.87	72.87	89.01	84.69	78.71	
Proposed			20,575,616	110	85.12	**76.23**	**78.60**	**78.60**	89.64	86.23	85.60	



Result

This chapter presents the results of the experiments and analyses the data collected during the study. It highlights key findings and visualisations to provide a clear understanding of the impact and significance of the research. As outlined in Chapter 4, LiDAR points were successfully projected onto the image plane using transformation matrices. This allowed the association of 2D image pixels with their corresponding 3D LiDAR depth values. The YOLOv5 model was then used to detect objects in the 2D images, and the depths of these objects were determined based on the projected LiDAR data. The integration of the two provided a clear understanding of the spatial relationships between detected objects and their depths. In LiDAR point cloud data, the ground plane refers to the flat surface detected by the LiDAR sensor, typically representing roads or other horizontal surfaces. By removing the ground plane from the data, the focus is shifted to relevant objects, such as vehicles, pedestrians, and obstacles. This step reduces computational load, minimises false positives, and enhances sensor fusion. With the ground plane removed, LiDAR points can be more accurately projected onto the image plane, allowing for precise depth association with image pixels. The YOLOv5 model successfully detects objects in the 2D image, and the depth information is determined using the LiDAR data. The integration of GPS and IMU data further enables the localisation of detected objects in global coordinates, providing a clear understanding of their positions in the real world. This comprehensive approach allows for accurate depth estimation and global localisation of detected objects, which is essential for tasks like autonomous navigation, advanced driver assistance systems, and 3D reconstruction. The ability to transform coordinates between different reference frames enhances situational awareness and improves operational efficiency. After detecting objects in the 2D image and estimating their depth from the LiDAR data, the results need to be validated for accuracy. As shown in Figure 4.3, the distance between the IMU and the Velodyne LiDAR is 0.81 meters, and the distance between the Velodyne and the Camera is 0.27 meters. The estimated depth from the camera space must account for these distances, and the total depth should be verified against the actual depth measured by the IMU to ensure accuracy.

CONCLUSION

This article introduces a one-stage 3D object detection framework that combines LiDAR and camera data, enhanced by three attention mechanisms to improve detection accuracy. First, the HA mechanism is applied to the input RGB images to generate RGBD images, which include depth information. Next, the GFA mechanism is used during feature extraction for both the RGB images and the Bird's Eye View (BEV) branches. This helps capture important features from both the image channels and spatial dimensions. Finally, the RA mechanism is used to fuse the paired view-specific regions of interest (ROIs) for better object detection. The proposed method significantly improves 3D object detection performance and surpasses other existing LiDAR and camera-based methods in accuracy. Looking ahead, the process of generating BEV images could potentially be replaced with a learnable feature generator, such as Second. Additionally, the method for anchor generation could shift from anchor-based to anchor-free, further improving performance. By leveraging both RGB images and LiDAR point clouds, 3D object detection using this combined approach could achieve even better results than methods relying solely on LiDAR.

REFERENCES

[1] Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., & Heide, F. (2020). Overcoming fog in autonomous driving using deep multimodal sensor fusion without directly seeing the fog. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[2] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). Muscones: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[3] Charvat, G. L. (2014). Small and Short-Range Radar Systems. Press.

[4] Chen, Y., Tai, L., Sun, K., & Li, M. Y. (2020). Monopair: monocular 3D object detection using pairwise spatial relationships. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., & Zhao, F. (2022). Graph-Detr3D: Rethinking overlapping regions for multi-view 3D object detection. In Proceedings of the ACM Multimedia Conference (ACMMM), 2022.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2020.

[7] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[8] Huang, J., & Huang, G. (2022). BEVDet4D: Exploiting temporal cues in multi-camera 3D object detection. arXiv preprint arXiv:2203.17054.

[9] Huang, J., & Huang, G. (2022). BEVPoolV2: A cutting-edge implementation of BEVDet for deployment. arXiv preprint arXiv:2211.17111.

[10] Huang, J., Huang, G., Zhu, Z., Ye, Y., & Du, D. (2021). BEVDet: High-performance multi-camera 3D object detection in bird's-eye view. arXiv preprint arXiv:2112.11790.

[11] Kim, Y., Choi, J. W., & Kum, D. (2020). GRIFNet: Gated region of interest fusion network for robust 3D object detection from radar point clouds and monocular images. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

[12] Kim, Y., Kim, S., Choi, J. W., & Kum, D. (2023). CRAFT: Camera-radar 3D object detection with spatio-contextual fusion transformer. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2023.

[13] Kim, Y., Shin, J., Kim, S., Lee, I.-J., Choi, J. W., & Kum, D. (2023). CRN: Camera-radar network for accurate, robust, and efficient 3D perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[14] Kingma, D. P. & Ba, J. (2015). Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

[15] Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Point Pillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[16] Li, P., Chen, X., & Shen, S. (2019). Stereo R-CNN-based 3D object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.