Enhanced Sparse Convolutional Detection Model for 3D Object Detection in Autonomous Vehicles Adapted to Traffic Conditions in Vietnam

Vu Hoang Dung*, Nguyen Trung Kien†, and Do Thanh Ha ‡
* Phenikka-X Company, Hanoi, Vietnam
E-mail: dungvh@phenikaa-x.com
† Hanoi University of Science and Technology
E-mail: kien.nt200303@sis.hust.edu.vn
† VNU University of Science
E-mail: dothanhha@hus.edu.vn

***Abstract*—The paper introduces a novel deep-learning model for enhancing 3D object detection in autonomous vehicles tailored to Vietnamese traffic conditions. The model improves efficiency by computing features only for relevant voxels, reducing computational costs and accelerating convergence through direct point cloud transformations. To address the limitations of the KITTI 3D dataset, a new proprietary dataset was created using Lidar16 and Lidar32 sensors, capturing specific Vietnamese traffic conditions. Experimental results on both the KITTI 3D and Phenikaa-X datasets show promising improvements in object detection performance, advancing techniques in 3D object detection and enhancing safety and reliability in diverse traffic environments.**

I. Introduction

In autonomous driving, 3D objects refer to entities detected and classified within three-dimensional space. These objects include vehicles, pedestrians, and static objects. Detecting and classifying these objects is crucial for the autonomous driving system to comprehend its surroundings and facilitate safe decision-making.

The computer vision approach for 3D object detection is mainly based on analyzing the image data resource or the characteristics of the image data. Various types of camera sensors usually capture image data. It comprises 2D images acquired from monocular cameras or 2D images with depth information obtained from stereo cameras. For image data in the form of 2D, the 3D object detection approaches analyze 2D images and then infer 3D information about objects [1][2]. Generally, the approach based solely on image data yields promising results but is constrained by the absence of depth information. Despite advancements in in-depth recovery from images, it remains an inherently unstable inverse problem. Variations in the 3D pose of the same object can result in distinct visual appearances on the image plane, complicating representation learning. Additionally, cameras are passive sensors, so images are susceptible to environmental factors such as varying lighting conditions (e.g., nighttime) or adverse weather like rain.

Point cloud data is gathered from 3D sensors like Lidar, capturing detailed 3D information about objects [3][4][5][6]. Point cloud data represents the surrounding 3D environment as a set of points, each with specific positional information in 3D space. This data aids in precisely determining the object's location in 3D space without additional transformations. However, detecting 3D objects from point clouds remains challenging due to several factors. Point clouds are often sparse, irregular, and unordered. Additionally, they may contain missing points or suffer from self-occlusion, further complicating accurate object detection data from images and point clouds are collected simultaneously from camera sensors and Lidar, providing comprehensive object characteristics (such as color) and precise object positions in 3D space based on point cloud information.

Processing combined data can help alleviate the limitations of each image data type [7]. However, establishing correspondence between the point cloud and the pixels presents a challenging problem that requires resolution. Thus, the prominent research direction receiving significant attention is enhancing point clouds-based approaches. One direction is representing the point cloud using voxels [8][9], which can improve resolution and accuracy but at the cost of increased computational complexity and memory usage. Another direction involves representing the point cloud as a point set, which can improve algorithm performance but add complexity to machine learning algorithms.

This paper researches and develops a 3D object detection method based on Lidar information. In more detail, we use an advanced fusion technique, which operates at the point level, and then develop a sophisticated sparse convolutional detection framework to enhance the performance of 3D detection in autonomous. The proposed approach improves detection performance, reduces computational complexity and inference latency, and offers a more efficient and reliable solution for autonomous driving applications. In addition in this paper, we also propose to enhance the KITTI 3D published dataset by overcoming its restricting 180-degree viewing angle and labeling the front portion within the camera's field of view to a comprehensive 360◦ view of the surrounding objects.

The rest of this paper is organized as follows: Section II provides a comprehensive review of related work and combined data from multiple sensors. Section III describes our proposed hybrid approach, including the detailed architecture and fusion technique employed to integrate LiDAR and camera data effectively. Section IV outlines the experimental setup and obtained results showcasing our method's performance improvements and advantages compared to baseline approaches. Finally, Section V concludes the paper with a summary of our contributions and critical outcomes.

II. Deep Learning Approaches in 3D Object Detection for Autonomous Driving

Over the past decade, we have witnessed significant advancements in 3D object recognition techniques, particularly those based on monochrome or dual-camera images [10], [11]. However, detecting 3D objects from monochrome images remains challenging, primarily due to the limited depth information available from a single image. Various methods, such as depth inference from images, leveraging geometric constraints, and shape assumption, have been proposed to tackle this issue, but they only offer partial solutions. One promising direction is a regression of 3D box parameters from images using a convolutional neural network (CNN). This approach, whether it focuses on predicting all 3D box parameters at once or initially predicts 2D boxes and then upgrades them to 3D boxes, holds the potential in 3D object detection.

3D object detection based on dual cameras is perspective direction since it uses two cameras positioned at a distance from each other to capture the same scene. While dual camera-based algorithms benefit from using geometric constraints to infer depth information, they also face limitations due to the need for precise calibration and synchronization.

Multi-view 3D detection is a direction that utilizes multiple cameras to capture the same scene from different angles. These images are then projected into a unified Bird's Eye View (BEV) space for processing by a BEV-based detector[12], [13]. The advantage of multi-view 3D detection is that it provides better detection than approaches relying on a single or dual camera, as multiple camera views offer more information about objects in the scene. However, multi-view 3D detection faces challenges in accurately transforming between camera views and BEV space without precise depth information, leading to misalignment between image pixels and their BEV positions. In addition, collecting and processing image data from multiple cameras can be costly and complex.

Unlike images, where pixels are evenly distributed on a plane, point clouds represent a sparse and irregular 3D data structure. Researchers have developed numerous 3D object detection models to explore pixel-level information based on point cloud data collected. The authors in Lidar [14],[15] proposed a method that uses a voxel grid to divide the point cloud into small cells, allowing for more efficient feature extraction.

Point-based 3D object detection relies on deep learning networks to detect objects within point clouds [16]–[18]. These methods, adapted from deep learning techniques used for images, involve two primary components: point cloud sampling and feature learning. Point cloud sampling, a critical step, involves selecting a subset of points from the original cloud using random, weighted, and structural sampling methods. This process proposes to reduce data size and enhance learning efficiency. Feature learning, the second component, entails extracting features from sampled points using point cloud operators. These features are then utilized for object classification within the point cloud.

To facilitate feature extraction, mesh-based 3D object detection involves converting point clouds into discrete grid representations, such as voxels, pillars, or top-view (BEV) maps [19]. Then, a deep learning model extracts features from these meshes. The methods in this direction involve two key components: grid-based representation and grid-based neural networks. Grid-based representation converts the point cloud into a discrete grid by dividing 3D space into small cells. Grid-based neural networks process these grid representations effectively in 3D space. Generally, mesh-based 3D object detection outperforms point-based approaches.

3D object detection using a combination of point and voxel data involves a hybrid architecture with two main frameworks: the one-stage and the two-stage. The one-stage framework directly processes the point cloud using points and voxels. Points are utilized for local feature extraction, while voxels contribute to global feature extraction. These extracted features are then employed for object detection within the point cloud. The two-stage framework operates in two distinct steps: the first one employs point- or grid-based methods for feature extraction, and the second one utilizes voxel-based methods specifically for object detection.

Another direction in 3D object detection involves integrating information from cameras and Lidar sensors, known as Early-Fusion-based 3D object detection [20]. Cameras can extract semantic features with color information, while Lidar excels in 3D positioning and structural information. The early fusion-based 3D object detection approach initiates with a 2D detection or segmentation network that extracts information from the image. This information is then transmitted to the Lidar point cloud. Depending on the approach to information fusion, early fusion-based 3D object detection methods are classified into region-level knowledge fusion and point-level knowledge fusion. Region-level knowledge fusion utilizes image-derived information to localize regions in the Lidar point cloud likely to contain objects. Point-level fusion methods enhance points in the Lidar point cloud using semantic features extracted from images. While Early-Fusion-based 3D object detection approaches show promise, they have the drawback of increased computational costs and inference time latency.

Late-Fusion-based 3D object detection approaches [16], [21], [22] leverage outputs from separate Lidar-based and camera-based object detectors. The Lidar detector provides 3D spatial information (position and size), while the camera detector contributes shape and color information. These methods

often have good results by capitalizing on the strengths of both approaches. However, the challenges are in fully integrating semantic information from both modalities, which can lead to inaccurate detection outcomes in specific scenarios.

III. Proposed Approaches for 3D Object Detection

This section explores the Sparse Convolutional Detection (SECOND) model, incorporating enhancement to conventional convolutional network architectures. One of the paper's contributions is adopting spatially sparse convolutional networks, which compute features only for voxels containing points, significantly reducing computational costs in the SECOND model. In addition, another contribution to adapting a dataset that can adjust to Vietnamese conditions traffic is presented in this section.

A. Enhanced Sparse Convolutional Detection Model

The architecture of the SECOND Network comprises the following components: Point Cloud, Voxel Feature Extractor, Sparse Convolution Layers, and Region Proposal Network (RPN). Point cloud grouping organizes point cloud data into coordinate tables, encoding voxel positions and actual point cloud data. The voxel feature extractor converts raw point clouds into voxel features, a crucial step in the process. Voxels, 3D cells that divide space into grids, are the backbone of the SECOND Network, which employs two Voxel Feature Extractor (VFE) layers to extract features from point clouds within voxels and encode them into voxel representations [14]. The sparse convolution layers, designed to operate on predominantly empty or sparse data spaces, such as those found in point cloud datasets, are the workhorses of the Network, extracting higher-level features from voxel representations, aiding in object identification within the point cloud. The region proposal network (RPN) generates bounding boxes around potential objects within the point cloud, a practical application of the Network's capabilities. These bounding boxes are subsequently utilized for object classification and localization, further demonstrating the real-world relevance of the SECOND Network.

In SECOND network architecture, computing sparse convolutions for feature extraction has significant challenges. In [23], the author uses regular sparse convolution and subarray sparse convolution (SSCN) for feature extraction. However, this approach has a limitation: it overlooks feature extraction from surrounding data and focuses only on activated inputs to produce activated outputs. Recognizing this, the paper [24] proposes a solution by modifying the model backbone. This modification involves integrating more targeted sparse convolution techniques, specifically focused sparse convolution (FSCN), to enhance the model's performance.

FSCN operates under the principle of fine-grained sparsity, where sparsity is distributed across the network. Fine-grained sparsity allows for more nuanced control over which weights contribute to the output, potentially leading to more efficient computations.

Conducting convolutions, mainly extracting voxel features, is the most time-intensive aspect of model execution. However, sparse convolutions significantly accelerate these calculations and processing, meeting practical demands. The paper investigates the potential of the SECOND model of replacing the backbone with FSCN to assess model accuracy and processing speed.

The total loss used during training enhanced SECOND model is a combination of the discussed loss functions:

$$L_{total} = \beta_1 L_{cls} + \beta_2 (L_{reg-\theta} + L_{reg-other}) + \beta_3 L_{dir} \quad (1)$$

in which L_{cls} is the classification loss, $L_{reg-other}$ is the regression loss for position and size, $L_{reg-\theta}$ is the angle loss, L_{dir} is the direction classification loss, β_1 = 1.0, β_2 = 2.0, β_3 = 0.2 are constant coefficients in this loss formula.

B. Constructing the dataset

The KITTI 3D dataset is a published dataset for computer vision in general and autonomous driving systems in particular. This dataset, derived from self-driving vehicles, includes images, point clouds, and associated information. However, when deploying the KITTI dataset in Vietnam presents several challenges. The first dataset reflects European (specifically German) road conditions characterized by well-demarcated roads conducive to object identification. The second one is the representation of vehicles. The KITTI dataset predominantly features cars, while Vietnamese roads are dominated by motorbikes and bicycles. This disparity can significantly impact the dataset's applicability in Vietnam. The third one is that the Lidar 64 sensor's point cloud resolution in the dataset yields high point density but requires a costly sensor. The last one is that the dataset labels only objects within the camera's visible range, resulting in partial vehicle annotations that enclose the front part of the vehicle with bounding boxes around the corresponding point clouds of identified objects.

Restricting the viewing angle is a notable limitation, as it only labels the front portion within the camera's field of view. This limitation impacts recognizing specific objects crucial for autonomous vehicles, necessitating a comprehensive 360° view of surrounding objects. To overcome this limitation, we proposed to enhance the KITTI 3D dataset's perspective as follows:

1. Utilize a deep learning model with pre-trained weights to re-identify all surrounding objects, extending beyond the camera's limited field of view. The focus is on achieving high accuracy rather than processing speed, aiming to reduce the need for post-calibration adjustments.
2. Employ the open-source SUSTECH tool [24], facilitating the visualization and information editing of 3D objects within a user interface. This tool aids in refining object-related data to enhance accuracy in the object recognition process.
3. Iteratively apply Steps 1 and 2 to expand the dataset's coverage and refine object labels, progressively enhancing the dataset's quality for subsequent training purposes.

To ensure a proposed model can work well in actual road conditions in Vietnam, we also train the model on a real dataset with 3500 kilometers of data from various rod types in Vietnam. The data is collected using Lidar16 and Lidar32 sensors since these sensors are more cost-effective and suitable for Vietnam's conditions. In addition, this dataset is standardized with labels following a uniform structure compatible with various published data sources. Finally, the data collection is aligned with the characteristics of Vietnam's roads, focusing on critical vehicles such as cars, motorbikes, bicycles, and pedestrians. Figure I presents the data using Lidar 16 and Lidar 32 in the street in Vietnam.

IV. Experiment Results

Besides using the published dataset Kitti3D, we also use the dataset described in Table II.

In Table II, Kit3D-1 and Kit3D-2 having 500 and 7,481 samples Kitti3D dataset, respectively. These samples were collected using a Lidar Velodyne 64 sensor and labeled for objects within a 180-degree frontal view of the vehicle. In addition, Kit3D-2 has added 500 individually-developed samples generated as described in Section III.B. Kit3D-3 is a dataset enhanced from Kit3D-2 with more than 1000 individually-developed samples and manually labeled within a 360-degree frontal view of the vehicle. It contains 1000 customized manually labeled samples, also collected using a Lidar Velodyne 64 sensor.

The CUS-1 and CUS-2 datasets also take 2000 and 1000 enhanced samples from Kitti3D with labels that cover a 360-degree view of the vehicle. Notably, the CUS-1 dataset has 1000 individually developed samples using Lidar 16 sensors. In addition, the CUS-2 dataset has 500 individually developed samples using Lidar 16 sensors and 1000 individually developed samples using Lidar 64 sensors. All samples in CUS-1 and CUS-2 are customized to cover a 360-degree view of the vehicle.

The performance of the deep-learning model is evaluated using AP and $AP|RN$ (with N = 40) measures. They are widely adopted in object recognition studies over 3D datasets. These metrics comprehensively evaluate object recognition model performance, encompassing accuracy and completeness. AP is typically employed to assess overall model performance, while $AP|RN$ (see Equation 2) helps evaluate model performance under conditions where high precision is crucial.

$$AP|_{R_N} = \frac{1}{N} \sum_{r \in R} P_{interpolare}(r), \tag{2}$$

Given $R = [r_0, r_0 + \frac{r_1 - r_0}{N-1}, r_0 + \frac{2(r_1 - r_0)}{N-1}, ..., r_1]$, the precision corresponding to each recall level r is interpolated. This interpolation is achieved by identifying the maximum precision value for which the recall value is greater than or equal to r, as expressed by Equation 3.

$$P_{interpolare}(r) = \max_{\tilde{r}:\tilde{r} \geq r} P(\tilde{r}) \tag{3}$$

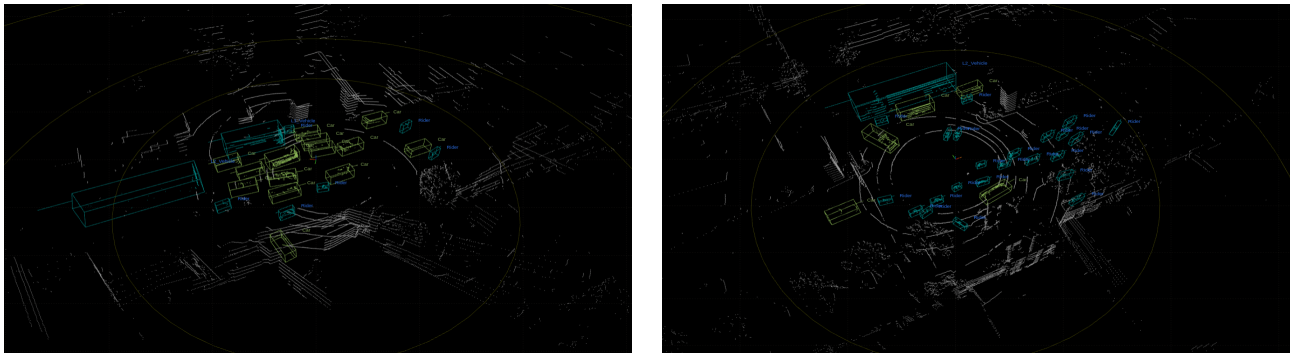
In the experiment, the rotated threshold IoU_{3D} is 0.7 for cars and 0.5 for pedestrians and cyclists. The IoU_{BEV} is computed by projecting the 3D bounding box onto the ground plane for BEV detection.

The first experiment evaluates the performance of using backbone FSCN instead of voxel backbone in 3D detection. The deep learning model with different backbones is trained on the KITTI dataset and then tested on Kit3D-1. Table **??** presents the obtained data and indicates that adopting the FSCN convolution backbone enhances model accuracy. However, utilizing the focal backbone prolongs model processing time since it takes 19.57s average to detect objects, while the voxel backbone uses 17.49s.

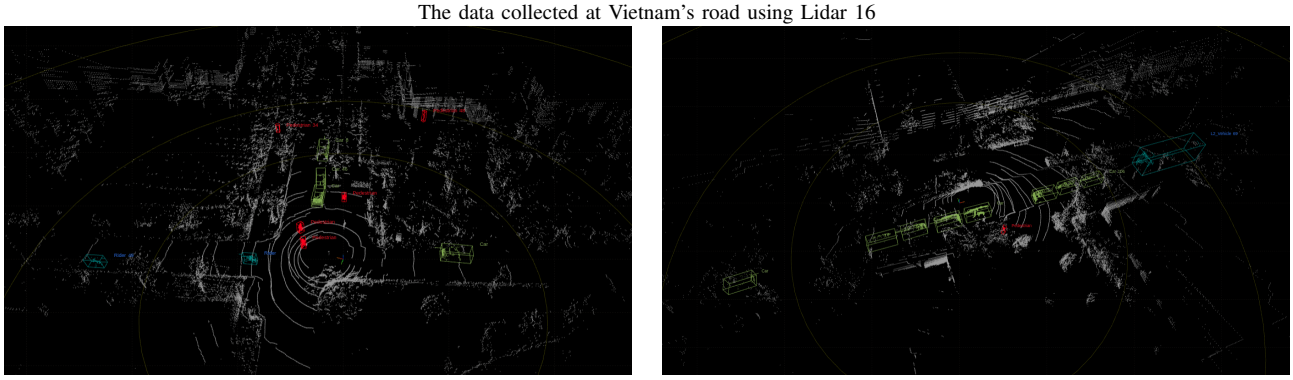
From the first empirical findings, we observe that modifying the model yields minimal changes. Therefore, we focus on leveraging diverse datasets in the following experiment to enhance the model's adaptability to Vietnamese traffic conditions. In more detail, we focus on data processing techniques as follows. First, training models on KIT3D-1, KIT3D-2, and KIT3D-3 datasets will be used, and then the performance on real datasets in Vietnamese traffic conditions will be verified. This real dataset belongs to the Phenikaa-X corporation. Due to confidentiality requirements, this dataset is not publicly accessible and is solely permitted to support the implementation of this paper. The dataset comprises samples collected using Lidar 16 and 32 sensors to capture 3000 kilometers of data from various road types in northern Vietnam, including highways, interprovincial roads, urban areas, and residential streets and 500 kilometers within Phenikaa University's campus and surrounding areas. This dataset is directly developing for Phenikaa's autonomous vehicles.

Table IV clearly illustrates the challenges we face. When the models are trained on the public dataset, they struggle to recognize 3D objects in Vietnamese traffic conditions, achieving only 0.94% accuracy with the Kit3D-1 dataset and 17.77 % with the Kit3D-2 training dataset and 18.43 % with the training dataset Kit3D-3. The low accuracy is a direct result of the limitations in the Kit3D-1 dataset, which only includes labels for the 180-degree frontal view of the vehicle. This results in the model's inability to identify surrounding objects (covering a 360-degree view of the vehicle). Moreover, as these datasets are public and not specifically tailored to Vietnamese environmental conditions, identifying accurate vehicle objects becomes challenging, thus leading to lower accuracy rates.

Table IV also indicates that the quality of data collection sensors significantly impacts dataset quality in terms of point cloud density and object distribution upon detection. It's crucial to emphasize that employing a combination of public and individually developed datasets is a step in the right direction to enhance model adaptability. However, these promising results are unsatisfactory due to these datasets' sparse point cloud density. Sparse point cloud density refers to a situation where the Lidar sensors have not captured enough data points in a given area, leading to gaps in the data. The Kit3D-3 dataset, despite being an extension of the public KITTI dataset that



The data collected at Vietnam's road using Lidar 16



The data collected at the campus of Phenikka University in Vietnam using Lidar 32

TABLE I
EXAMPLE COLLECTED DATA USING LIDAR 16 AND LIDAR 32, RESPECTIVELY IN VIETNAM

	KITTI dataset			**mixed dataset**	
	Kit3D-1	**Kit3D-2**	**Kit3D-3**	**CUS-1**	**CUS-2**
Lidar	64	64	64	64 + 16	64 + 32 + 16
Angle	180^0	180^0 + 360^0	360^0	360^0	360^0
Number of samples	500	7481 + 500	7481 + 1500	2000 + 1000	1000 + 1000 + 500

TABLE II
ALL DATASET IN THE PAPER

Measure	**Voxel**	**Voxel Focal**
AP	81.765 %	82.395 %
APR40	84.1 %	84.29 %

TABLE III
PERFORMANCE OF THE SECOND MODEL WITH DIFFERENT BACKBONE.

Measure	**KIT3D-1**	**KIT3D-2**	**KIT3D-3**	**CUS-1**	**CUS-2**
AP	0.9	9.09	9.09	28.64	40.09
APR40	0.94	17.11	18.43	48.72	47.77

TABLE IV
PERFORMANCE OF THE MODEL TRAINED ON DIFFERENT DATASETS AND VERIFYING ON PHENIKAA'S DATASET.

was individually developed for improved accuracy, still falls short of practical deployment requirements due to the sparsely distributed point cloud density resulting from using Lidar16 sensors for data collection. This underscores the pressing need to enhance data quality, even when collected under Vietnam's traffic conditions, and emphasizes the necessity for sensor improvement.

The model trained on the CUS-1 and CUS-2 datasets achieved accuracies of 48.72% and 47.77%, respectively. These training models combine public data with individually developed samples, addressing limitations present in both datasets, such as unsuitable environmental conditions for Vietnamese traffic and insufficient point cloud density for effective object recognition. However, it's crucial to stress the need for continuous improvement. Our work is not done yet. Training additional datasets under Vietnamese environmental conditions using advanced Lidar sensors is recommended to improve model accuracy further and ensure the safer deployment of autonomous vehicles in traffic scenarios. This emphasis on continuous improvement underscores the importance of our collective efforts in this field.

V. CONCLUSION

This paper presents a deep learning-based model designed to enhance 3D object detection capabilities in autonomous ve-

hicles operating within the unique traffic conditions prevalent in Vietnam. By focusing on the computation of features for only the most relevant voxels, the model significantly reduces computational overhead and accelerates the convergence of the detection process through direct point cloud transformations. Addressing the shortcomings of the widely used KITTI 3D dataset, we evaluated our approach over an innovative proprietary dataset (Phenikaa-X datasets) utilizing Lidar16 and Lidar32 sensors to capture the nuances of Vietnamese traffic scenarios accurately. The experimental validation of the model across both the KITTI 3D and Phenikaa-X datasets demonstrates substantial advancements in object detection performance. These findings underscore the potential of the proposed model to significantly improve safety and reliability in various traffic environments, particularly in regions characterized by complex driving conditions, such as Vietnam.

ACKNOWLEDGMENT

This research was funded by the research project QG.23.71 of Vietnam National University, Hanoi.

REFERENCES

[1] S. Song and J. Xiao, *Deep sliding shapes for amodal 3d object detection in rgb-d images*, 2016. eprint: 1511.02300.

[2] X. Shen and I. Stamos, *Frustum voxnet for 3d object detection from rgb-d or depth images*, 2023.

[3] H. Ran, J. Liu, and C. Wang, "Surface representation for point clouds," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 920–18 930.

[4] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, "A hierarchical graph network for 3d object detection on point clouds," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 389–398.

[5] B. Cheng, L. Sheng, S. Shi, M. Yang, and D. Xu, "Back-tracing representative points for voting-based 3d object detection in point clouds," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8959–8968.

[6] P.-S. Wang, "Octformer: Octree-based transformers for 3d point clouds," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–11, Jul. 2023, ISSN: 1557-7368.

[7] Y. Zhang, J. Yu, X. Huang, W. Zhou, and J. Hou, *Pcr-cg: Point cloud registration via deep explicit color and geometry*, 2023.

[8] X. Du, M. H. A. J. au2, S. Karaman, and D. Rus, *A general pipeline for 3d detection of vehicles*, 2018. eprint: 1803.00387.

[9] V. A. Sindagi, Y. Zhou, and O. Tuzel, *Mvx-net: Multimodal voxelnet for 3d object detection*, 2019. eprint: 1904.01649.

[10] F. Ma and S. Karaman, *Sparse-to-dense: Depth prediction from sparse depth samples and a single image*, 2018. eprint: 1709.07492.

[11] X. Cheng, P. Wang, C. Guan, and R. Yang, *Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion*, 2019. eprint: 1911.05377.

[12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6526–6534.

[13] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1–6.

[14] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, *Voxel r-cnn: Towards high performance voxel-based 3d object detection*, 2021. eprint: 2012.15712.

[15] Y. Zhang, W. Feng, Y. Quan, G. Ye, and G. Dauphin, "Dynamic spatial-spectral feature optimization-based point cloud classification," *Remote. Sens.*, vol. 16, p. 575, 2024.

[16] Z. Yin, H. Sun, N. Liu, H. Zhou, and J. Shen, *Fgfusion: Fine-grained lidar-camera fusion for 3d object detection*, 2023. eprint: 2309.11804.

[17] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[18] L. Ma, Y. Li, J. Li, W. Tan, Y. Yu, and M. A. Chapman, "Multi-scale point-wise convolutional neural networks for 3d object segmentation from lidar point clouds in large-scale environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 821–836, 2021.

[19] Y. Cui, R. Chen, W. Chu, *et al.*, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, Feb. 2022, ISSN: 1558-0016.

[20] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-aware unsupervised deep lidar-stereo fusion," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6332–6341.

[21] T. He, P. Sun, Z. Leng, C. Liu, D. Anguelov, and M. Tan, *Lef: Late-to-early temporal fusion for lidar 3d object detection*, 2023. eprint: 2309.16870.

[22] Y. Li, A. W. Yu, T. Meng, *et al.*, *Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection*, 2022. eprint: 2203.08195.

[23] Y. Yan, Y. Mao, and B. Li, "SECOND: sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[24] E. Li, S. Wang, C. Li, D. Li, X. Wu, and Q. Hao, "Sustech points: A portable 3d point cloud interactive annotation platform system," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1108–1115.