Towards broader spatial-context 3D object detection for autonomous driving

Álvaro Ramajo-Ballester
Intelligent Systems Lab
Universidad Carlos III de Madrid
Madrid, Spain
aramajo@ing.uc3m.es
0000-0001-9425-9408

Arturo de la Escalera Hueso
Intelligent Systems Lab
Universidad Carlos III de Madrid
Madrid, Spain
escalera@ing.uc3m.es
0000-0002-2618-857X

José María Armingol Moreno
Intelligent Systems Lab
Universidad Carlos III de Madrid
Madrid, Spain
armingol@ing.uc3m.es
0000-0002-3353-9956

***Abstract*—This work presents an exhaustive analysis and a quantitative performance comparison between the use of information from infrastructure and vehicle mounted sensors for 3D object detection in autonomous driving environments. To do so, LiDAR point clouds have been used as the main data input and the most popular and well-established models have been considered for this task: Second, PointPillars and PV-RCNN. They have all been trained on the DAIR-V2X cooperative dataset, since it offers both the infrastructure and vehicle perspective. The broader spatial context and greater field of vision from an elevated point of view demonstrate superior performance by mitigating occlusions and overcoming the inherent limitations of a reduced perception range from onboard a vehicle. However, this comes with its own challenges to avoid losing detection capabilities for smaller objects. The main objective of this work is to provide a like-for-like comparison of the real performance difference, isolating the point of view as the only modified variable.**

***Index Terms*—3D object detection, LiDAR-based object detection, autonomous driving**

I. Introduction

Autonomous driving has witnessed substantial advancements, seeking to facilitate vehicle navigation with minimal human intervention. This rapid development was aided by improvements in computer vision thanks to Deep Learning techniques, with the corresponding performance refinement in other related sub-fields such as visual vehicle identification, license plate recognition [1] and many others.

A pivotal facet of these systems is environmental perception, involving the interpretation of multi-modal data, including camera images and LiDAR-generated point clouds, to discern the geometry and semantic attributes of objects on the road. As the field of computer vision evolves, so do 3D object detection algorithms, marked by diverse methodologies and evaluation metrics.

A recent trend in this domain explores a novel research perspective: the detection of 3D objects from an elevated infrastructure vantage point. Positioned several meters above the ground, sensors in this configuration substantially broaden the field of view, mitigating occlusions between road elements and the surrounding environment. However, this approach introduces challenges, including the establishment of effective communication between vehicles and elevated sensors, protocol standardization across different devices, and other pertinent considerations. In addition, and focusing on the perception task, which constitutes the core of this work, increasing the range comes at a cost of a quadratic growth of search space, which can pose a problem for detecting smaller objects.

In order to offer a better understanding of the real performance difference across the most used models in the state of the art and points of view, this work provides a qualitative contrast between them. After training three selected models in each of the subsets (infrastructure and vehicle), it can be seen how each point of view comes with its own advantages and disadvantages.

II. Background

A. Foundations

The core objective of 3D object detection is to predict the characteristics of objects in three-dimensional driving scenarios based on sensory inputs. In this context, let $\mathcal{X}$ denote the input data, which could be LiDAR or RGB images, and $\mathcal{F}$ represent a detector parameterized by θ. The general formulation for 3D object detection can be expressed as:

$$\mathcal{B} = \mathcal{F}(\mathcal{X}; \theta) \tag{1}$$

where $\mathcal{B} = B_1, B_2, ..., B_n$ is a set of n 3D objects within a scene. The representation of a 3D object is crucial in this task, influencing the data required for subsequent prediction and planning processes. Typically, a 3D object is depicted as a cuboid, which encapsulates the object. This enclosure can be delineated by its 8 corners [2], 4 corners and heights [3], or more commonly, by the 7 parameters defining an oriented bounding box [4], [5] in equation (2).

$$\mathcal{B} = \{x, y, z, l, w, h, \theta, c\} \tag{2}$$

In this representation, (x, y, z) denotes the center coordinates of the cuboid; (l, w, h) denote its length, width, and height, respectively; θ is the yaw angle in the ground plane, and c indicates the corresponding class, such as car, pedestrian, etc.

B. Sensors

For 3D object detection, a diverse range of sensors can supply raw data, with cameras and LiDAR (Light Detection and Ranging) sensors emerging as the most commonly employed types.

Cameras offer an economical and readily available means of capturing scene details from specific perspectives. They generate images $\mathcal{X}_{cam} = \mathbb{R}^{W \times H \times 3}$, where W and H denote the width and height of the image, respectively, and each pixel comprises 3 RGB channels. Despite their cost-effectiveness, cameras face inherent limitations in 3D object detection. Primarily, they capture visual data exclusively, lacking the capacity to perceive the three-dimensional structure of a scene. To address this limitation, stereo cameras employ matching algorithms, aligning the correspondences in both left and right images to recover depth [6].

In contrast, LiDAR sensors facilitate the acquisition of detailed 3D scene structures by emitting numerous laser beams and measuring their reflective information. A LiDAR sensor produces a range image $\mathcal{X}_{lid} \in \mathbb{R}^{m \times n \times 3}$ using m beams and n readings in a single scan cycle. Each pixel in this range image encompasses 3 channels, corresponding to range, azimuth, and inclination in the spherical coordinate system. Converting spherical coordinates into Cartesian coordinates allows the transformation of range images into point clouds.

C. Datasets

As the data-driven era progresses, the accessibility of large-scale datasets has been continuously enriching and encouraging the community. Some of the most notable and publicly accessible datasets related to autonomous driving have been included. KITTI [7], [8], Waymo Open [9] and nuScenes [10] datasets stand out as the most popular ones, among others. An example of the last two is shown in fig. 1 and a more detailed comparison is presented in table I.

KITTI Dataset. Pioneering the domain of 3D object detection, the KITTI dataset was released in 2012. It contains LiDAR and visual data from 15,000 frames, with over 200,000 3D annotations across eight classes (car, van, truck, pedestrian, person, cyclist, tram and misc). For online scoreboard evaluation, only car, pedestrian, and cyclist labels are considered. Three difficulty levels (Easy, Moderate, and Hard) are defined based on the height of 2D bounding boxes, occlusion levels, and truncation.

Waymo Open. This dataset features annotations for 12 million 3D bounding boxes within 200,000 frames, distributed across 1,150 sequences. It spans four classes, with only three aligning with KITTI classes.

nuScenes. Manually labeling 1.4 million boxes across 40,000 frames, the nuScenes dataset comprises 1,000 sequences spanning 23 classes. However, only 10 of these classes are considered for detection.

It is important to note that testing labels are not available for these datasets. Researchers must submit their predictions to an online leaderboard server for assessment on the test set. Notably, nuScenes and Waymo Open capture data under various weather and lighting conditions, including rain, fog, snow, daytime and nighttime, unlike KITTI, which focuses solely on sunny days.

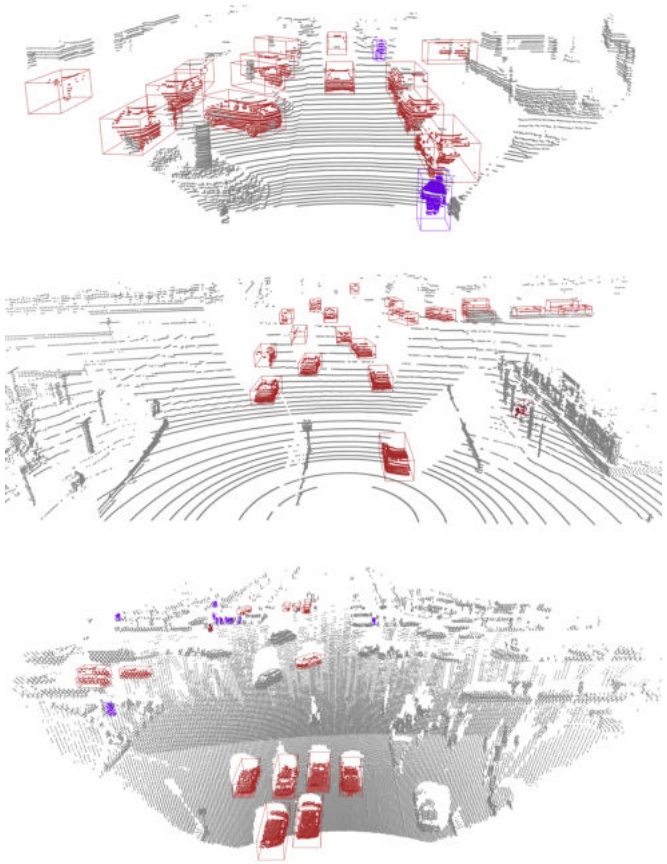

Fig. 1: KITTI (above), DAIR-V2X Coop. Vehicle (middle) and DAIR-V2X Coop. Infrastructure (below) datasets examples

In recent years, several dataset publications have been released, featuring infrastructure-mounted sensors, such as Rope3D [11], A9 Dataset [12], IPS300+ [13], and DAIR-V2X [14].

DAIR-V2X. One of the most complete, this dataset offers three versions: Vehicle Dataset (DAIR-V2X-V), comprising 22,325 LiDAR and image frames; Infrastructure Dataset (DAIR-V2X-I), including 10,084 point cloud and image frames, and Cooperative Dataset (DAIR-V2X-C), with 18,330 data frames from infrastructure and 20,515 from the vehicle for Vehicle-Infrastructure Cooperative (VIC) 3D object detection. The latter version was used in this study to compare model performances.

D. Evaluation metrics

The main evaluation measure for 3D object detection is Average Precision (AP), which is based on the same principles as its 2D equivalents [31]. Prior to delving into the nuanced similarities and distinctions of dataset-specific AP used in popular benchmarks, the basic definition of the AP metric is outlined as follows:

TABLE I: Comparison between publicly available datasets for 3D object detection, sorted by year [15]

Dataset	LiDAR	Images	3D annot.	Cl.	Night/Rain	View
KITTI [7]	15k	15k	200k	8	No/No	Onboard
Ko-FAS [16]	39k	19.4k	-	-	-/-	Infrastructure
KAIST [17]	8.9k	8.9k	-	3	Yes/No	Onboard
ApolloScape [18]	20k	144k	475k	6	-/-	Onboard
H3D [19]	27k	83k	1.1M	8	No/No	Onboard
Lyft L5 [20]	46k	323k	1.3M	9	No/No	Onboard
Argoverse [21]	44k	490k	993k	15	Yes/Yes	Onboard
A*3D [22]	39k	39k	230k	7	Yes/Yes	Onboard
A2D2 [23]	12.5k	41.3k	-	14	-/-	Onboard
nuScenes [10]	400k	1.4M	1.4M	23	Yes/Yes	Onboard
Waymo Open [9]	230k	1M	12M	4	Yes/Yes	Onboard
AIODrive [24]	250k	250k	26M	-	Yes/Yes	Virtual onboard
BAAI-VANJEE [25]	2.5k	5k	74k	12	Yes/Yes	Infrastructure
PandaSet [26]	8.2k	49k	1.3M	28	Yes/Yes	Onboard
KITTI-360 [27]	80k	300k	68k	37	-/-	Onboard
Argoverse 2 Sensor [28]	150k	1M	-	30	Yes/Yes	Onboard
ONCE [29]	1M	7M	417k	5	Yes/Yes	Onboard
Cirrus [30]	6.2k	6.2k	-	8	-/-	Onboard
Rope3D [11]	-	50k	1.5M	12	Yes/Yes	Infrastructure
A9 Dataset [12]	1.7k	5.4k	215k	8	Yes/Yes	Infrastructure
IPS300+ [13]	28k	57k	4.5M	7	Yes/-	Infrastructure
DAIR-V2X [14]	71k	71k	1.2M	10	-/-	Onboard / Infrast.

$$AP = \int_0^1 \max\{P(r'|r' \geq r)\}dr \tag{3}$$

where $P(r)$ represents the precision-recall curve; r' denotes each potential recall value, and r is the recall variable used in the integral calculation. When assessing precision and recall, the key difference from the 2D AP metric lies in the criteria for matching predictions with ground truths. KITTI introduces two widely adopted AP metrics: 3D AP and BEV AP. 3D AP aligns predicted objects with their corresponding ground truths if the 3D Intersection over Union (3D IoU) of two cuboids exceeds a specified threshold, while BEV AP is determined based on the IoU of two cuboids in the bird's-eye view (BEV IoU).

Calculating this area precisely is a challenging task. PASCAL VOC [31] introduced a different measure called interpolated AP_{R_N}. This metric calculates the average precision at various recall levels (R) evenly distributed across N levels, ranging from $r_0 = 0$ to $r_1 = 1$:

$$AP|_{R_N} = \frac{1}{N} \sum_{r \in R} P(r) \tag{4}$$

where $R = [r_0, r_0 + \frac{r_1 - r_0}{N-1}, r_0 + \frac{2(r_1 - r_0)}{N-1}, \ldots, r_1]$ and $P(r) = \max_{r':r' \geq r} P(r')$.

KITTI Benchmark [7]. Initially, the KITTI benchmark employed the interpolated $AP|R11$ metric before changing to $AP|R40$, following the suggestions from [32]. This adjustment aimed to facilitate a more fair comparison of scores. KITTI maintains distinct leaderboards for 3D object detection and Bird's Eye View (BEV) detection tasks. As previously said, it differentiates three levels of difficulty: easy, moderate and hard, regarding the occlusion and height of the bounding boxes:

- Easy: minimum bounding box height, 40 pixels; maximum occlusion level, fully visible; and maximum truncation, 15 %
- Moderate: minimum bounding box height, 25 pixels; maximum occlusion level, partly occluded; and maximum truncation, 30 %
- Hard: minimum bounding box height, 25 pixels; maximum occlusion level, difficult to see; and maximum truncation, 50 %

The 3D object detection and orientation estimation benchmark is divided into four parts. Firstly, the 2D AP, which is the classical 2D object detection with average precision (AP) metric. True positives should overlap by more than 50% and multiple detections of the same object are considered false positives. In addition, the performance of jointly detecting objects and estimating their 3D orientation is calculated with the average orientation similarity (AOS) as described in [7], giving the 3D AP. In the third place, a classification and continuous orientation regression for orientation similarity and, finally, the BEV (Bird's-Eye View) AP.

Waymo Benchmark [9]. In a similar vein, Waymo Benchmark introduces interpolated AP_{R21} and Average Precision weighted by heading (APH) metrics. Waymo evaluates performance across 21 evenly spaced recall levels ($r_0 = 0, r_1 = 1, N = 21$), with IOU thresholds set at 0.7 for vehicles and 0.5 for pedestrians and cyclists. For APH, the true positives are weighted by the heading accuracy:

$$w_h = \min(|\hat{\theta} - \theta|, 2\pi - |\hat{\theta} - \theta|)/\pi \quad (5)$$

where $\hat{\theta}$ and θ are the predicted and ground truth azimuth, respectively. Two levels of difficulty are included in the benchmark: L1 for bounding boxes with more than five lidar points and L2 for those with between one and five points.

nuScenes Benchmark [10]. The nuScenes benchmark employs a custom metric known as NuScenes Detection Score (NDS). To calculate this score, a set of error metrics is defined to measure translation (ATE), scale (ASE), orientation (AOE), velocity (AVE), and attribute errors (AAE), all with a 2-meter center distance threshold. These errors (ϵ) are then transformed into scores using the formula depicted in equation (6).

$$s_i = \max(1 - \epsilon_i, 0) \quad (6)$$

These metrics are weighted afterwards to calculate the NDS:

$$NDS = \frac{1}{10}\left[5mAP + \sum_i s_i\right] \quad (7)$$

where mAP is calculated by a BEV center distance with thresholds 0.5m, 1m, 2m, 4m [33].

Since this paper aims to establish a quantitative comparison between the performance with sensors mounted on the vehicle and on the infrastructure, the KITTI benchmark was used.

E. Model architectures

In this work, 3 different models have been adapted to the DAIR-V2X cooperative dataset:

- **Second**. The sensor processes an initial point cloud, transforming it into voxel features and coordinates, then employs two VFE (voxel feature encoding) layers along with a linear layer. After that, a sparse CNN is applied, followed by an RPN for generating the detection [34], as it can be seen in figure 2.

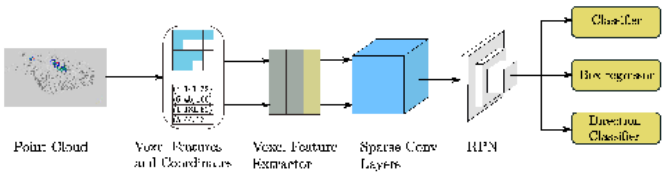

Fig. 2: Second architecture [34]

- **PointPillars**. The fundamentals of the network are a Pillar Feature Network, Backbone, and SSD Detection Head (figure 3). The unprocessed point cloud is converted to a stacked pillar tensor and pillar index tensor, which is forwarded to the encoder to learn the features that are then scattered back to a 2D pseudo-image for a convolutional neural network. After passing through the backbone, the final features are used by the detection head to predict 3D bounding boxes for objects [4].
- **PV-RCNN**. It involves a 3D voxel CNN that serves as the core for effective feature encoding and proposal generation using sparse convolution. To enhance the pooling of features from each 3D object proposal within the scene, two innovative operations are introduced: the voxel-to-keypoint scene encoding, which condenses all voxel information from the scene into a few feature keypoints, and the point-to-grid RoI feature abstraction, which efficiently combines scene keypoint features into RoI grids for predicting proposal confidence and refining location [35]. This architecture is presented in figure 4.

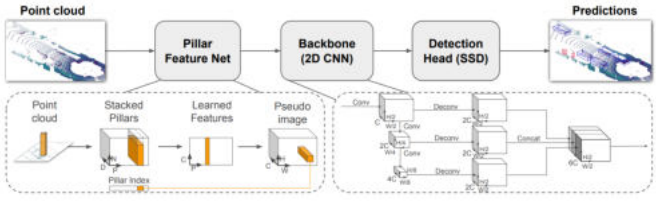

Fig. 3: PointPillars architecture [35]

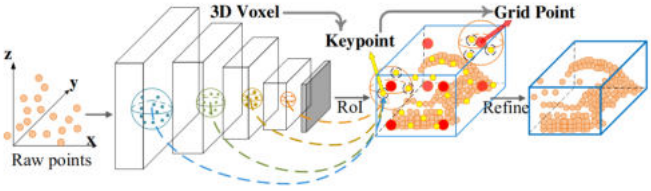

Fig. 4: PV-RCNN architecture [35]

F. Spatial context in 3D object detection

In the matter of 3D object detection evaluation, there are several metrics that provide specific aspects of the overall performance. The most common ones are 3D average precision and Bird's Eye View. In both of them, for a detection to be considered valid, the Intersection over Union of the predicted 3D or 2D bounding box has to be, at least, greater or equal to a certain threshold.

To have a more flexible estimation of the algorithm capabilities, two different threshold levels have been used: strict and loose. For the biggest objects of the considered classes –cars, in this work–, a strict IoU of 0.7 and a loose one of 0.5 have been selected. For smaller objects, cyclists and pedestrians, a strict IoU of 0.5 and a loose one of 0.25.

III. Experiments

The previous models have been adapted and fine-tuned for the two perspectives of the DAIR-V2X cooperative dataset: infrastructure and vehicle. All the trainings have been performed on 80 epochs although the specific configurations have varied across the different models. The data augmentation techniques include random horizontal flip (0.5 probability), rotation and scale transformations.

The performance difference in favor of the infrastructure perspective can be perceived, even though the point-cloud range was extended from 70 meters maximum in vehicle set to 100 meters maximum in infrastructure set.

TABLE II: 3D AP_{40} metric comparison

3D AP_{40}	Car@0.7	Car@0.5	Cyclist@0.5	Cyclist@0.25	Pedestrian@0.5	Pedestrian@0.25
Second (Infrastructure)	29.74	37.96	14.49	25.39	11.20	22.23
PointPillars (Infrastructure)	**38.10**	**42.02**	**25.22**	**27.85**	18.23	24.65
PV-RCNN (Infrastructure)	33.78	40.97	11.41	21.40	13.41	24.87
Second (Vehicle)	27.02	29.41	19.44	23.18	**20.86**	**32.14**
PointPillars (Vehicle)	35.05	34.80	24.91	27.32	20.79	27.99
PV-RCNN (Vehicle)	30.80	34.32	11.10	20.90	15.96	28.43

TABLE III: BEV AP_{40} metric comparison

BEV AP_{40}	Car@0.7	Car@0.5	Cyclist@0.5	Cyclist@0.25	Pedestrian@0.5	Pedestrian@0.25
Second (Infrastructure)	32.59	38.31	17.07	25.69	13.64	22.50
PointPillars (Infrastructure)	**41.55**	42.09	25.49	27.87	19.04	24.85
PV-RCNN (Infrastructure)	37.37	**43.18**	12.37	21.56	15.82	24.91
Second (Vehicle)	27.45	31.63	21.50	24.58	**24.64**	**32.68**
PointPillars (Vehicle)	33.62	36.41	**28.15**	**28.93**	23.99	28.13
PV-RCNN (Vehicle)	29.03	36.95	14.99	22.71	21.20	27.91

IV. Results

After setting the trainings as previously explained, the results of the experiments are exhibited in this section.

To convey a better and more comprehensive performance description, table II presents the 3D metrics calculated with a strict and loose IoU for each class. This threshold difference is especially important in the smallest classes, where the precision has more variance.

Table III shows the Bird's Eye View metric for each class and IoU level. Both tables show the average precision for the moderate difficulty on the validation set.

PointPillars generally outperforms other models across both 3D and BEV benchmarks, especially in car and cyclist detection. Second model exhibits competitive performance, particularly in pedestrian detection in vehicle perspective. PV-RCNN consistently lags behind PointPillars and Second models in most scenarios, indicating room for improvement in performance.

Furthermore, the infrastructure perspective is preferred for scenarios where comprehensive coverage and visibility are critical, especially for larger objects like cars and cyclists. It offers advantages in scenarios with fewer occlusions from infrastructure elements and other vehicles. In the other hand, vehicle perspective may be advantageous for scenarios where detection of smaller objects or detailed observations of nearby vehicles are paramount. It is better suited for scenarios with high vehicle density or complex traffic situations, where proximity to objects provides a better context.

V. Conclusions

According to the result tables, it can be deduced that a superior and oblique perspective from the sensors mounted on the infrastructure gives a performance increase when it comes to 3D vehicle detection. This can be explained due to the lesser number of occlusions and a clearer view.

However, the detection of smaller objects, such as cyclists and pedestrians, does not benefit from the field of vision of the infrastructure. In this cases, the greater distance to the LiDAR sensors causes to have less laser beams reflected onto the objects and, therefore, a less dense point cloud mesh for the models to extract information from. In addition, the greater distance range that was inputted to the model during training can have effects similar to the *curse of dimensionality* [36], as the objects become proportionally smaller.

To overcome these disadvantages, attention-based models [37] can be used to guide neural networks to better understand the underlying patterns for 3D object detection.

VI. Acknowledgments

Grants PID2021-124335OB-C21, PID2022-140554OB-C32 and PDC2022-133684-C31 funded by MCIN/AEI/10.13039/501100011033.

References

[1] Á. Ramajo-Ballester, J. M. Armingol Moreno, and A. de la Escalera Hueso, "Dual license plate recognition and visual features encoding for vehicle identification," *Robotics and Autonomous Systems*, vol. 172, p. 104608, 2024.

[2] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[3] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[4] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[5] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[6] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[10] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[11] X. Ye, M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, and E. Ding, "Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 341–21 350.

[12] C. Creß, W. Zimmer, L. Strand, M. Fortkord, S. Dai, V. Lakshminarasimhan, and A. Knoll, "A9-dataset: Multi-sensor infrastructure-based dataset for mobility research," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 965–970.

[13] H. Wang, X. Zhang, Z. Li, J. Li, K. Wang, Z. Lei, and R. Haibing, "Ips300+: a challenging multi-modal data sets for intersection perception system," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2539–2545.

[14] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[15] Á. Ramajo-Ballester, A. de la Escalera Hueso, and J. M. Armingol Moreno, "3D Object Detection for Autonomous Driving: A Practical Survey," in *9th International Conference on Vehicle Technology and Intelligent Transport Systems*, 2023, pp. 64–73.

[16] E. Strigel, D. Meissner, F. Seeliger, B. Wilking, and K. Dietmayer, "The ko-per intersection laserscanner and video dataset," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 1900–1901.

[17] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[18] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.

[19] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9552–9557.

[20] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Level 5 perception dataset 2020," https://level-5.global/level5/data/, 2019.

[21] Y. Chai, P. Sun, J. Ngiam, W. Wang, B. Caine, V. Vasudevan, X. Zhang, and D. Anguelov, "To the point: Efficient 3d object detection in the range image with graph convolution kernels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 000–16 009.

[22] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A* 3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2267–2273.

[23] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn *et al.*, "A2d2: Audi autonomous driving dataset," *arXiv preprint arXiv:2004.06320*, 2020.

[24] X. Weng, Y. Man, D. Cheng, J. Park, M. O'Toole, K. Kitani, J. Wang, and D. Held, "All-in-one drive: A large-scale comprehensive perception dataset with high-density long-range point clouds," *arXiv*, 2020.

[25] D. Yongqiang, W. Dengjiang, C. Gang, M. Bing, G. Xijia, W. Yajun, L. Jianchao, F. Yanming, and L. Juanjuan, "Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china," *arXiv preprint arXiv:2105.14370*, 2021.

[26] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang *et al.*, "Pandaset: Advanced sensor suite dataset for autonomous driving," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3095–3101.

[27] Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[28] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[29] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li *et al.*, "One million scenes for autonomous driving: Once dataset," *arXiv preprint arXiv:2106.11037*, 2021.

[30] Z. Wang, S. Ding, Y. Li, J. Fenn, S. Roychowdhury, A. Wallin, L. Martin, S. Ryvola, G. Sapiro, and Q. Qiu, "Cirrus: A long-range bi-pattern lidar dataset," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5744–5750.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[32] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.

[33] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: a survey," *Pattern Recognition*, p. 108796, 2022.

[34] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely Embedded Convolutional Detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.

[35] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection," Apr. 2021.

[36] R. E. Bellman, *Dynamic Programming*. Courier Corporation, Jan. 2003.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.