Can Existing 3D Monocular Object Detection Methods Work in Roadside Contexts? A Reproducibility Study

Silvio Barra[2], Mirko Marras[1(✉)], Sondos Mohamed[1], Alessandro Sebastian Podda[1], and Roberto Saia[1]

[1] University of Cagliari, Cagliari, Italy
{mirko.marras,sondoswa.mohamed,sebastianpodda,roberto.saia}@unica.it
[2] University of Naples "Federico II", Naples, Italy
silvio.barra@unina.it

Abstract. Detecting 3D objects in images from urban monocular cameras is essential to enable intelligent monitoring applications for local municipalities decision-support systems. However, existing detection methods in this domain are mainly focused on autonomous driving and limited to frontal views from sensors mounted on the vehicle. In contrast, to monitor urban areas, local municipalities rely on streams collected from fixed cameras, especially in intersections and particularly dangerous areas. Such streams represent a rich source of data for applications focused on traffic patterns, road conditions, and potential hazards. In this paper, given the lack of availability of large-scale datasets of images from roadside cameras, and the time-consuming process of generating real labelled data, we first proposed a synthetic dataset using the CARLA simulator, which makes dataset creation efficient yet acceptable. The dataset consists of 7,481 development images and 7,518 test images. Then, we reproduced state-of-the-art models for monocular 3D object detection proven to work well in autonomous driving (e.g., M3DRPN, Monodle, SMOKE, and Kinematic) and tested them on the newly generated dataset. Our results show that our dataset can serve as a reference for future experiments and that state-of-the-art models from the autonomous driving domain do not always generalize well to monocular roadside camera images. Source code and data are available at https://bit.ly/monocular-3d-odt.

Keywords: Dataset · Object Detection · 3D Vision · Roadside Camera

1 Introduction

Monocular cameras play an important role in urban areas, in which they are commonly used in intersections and other high-risk locations to capture valuable data. Detecting objects in images from monocular cameras is critical for

R. Basili et al. (Eds.): AIxIA 2023, LNAI 14318, pp. 321–335, 2023.
https://doi.org/10.1007/978-3-031-47546-7_22

developing intelligent monitoring applications that can assist local municipalities in making timely and informed decisions [1–4,15,16]. With such applications, governments can obtain real-time and accurate information regarding traffic patterns, road conditions, and potential hazards. Differently from traditional 2D object detection methods [8,18,21–24,30,31,42], applying 3D object detection approaches offer significant advantages. By providing a more complete understanding of the scene and enabling the detection of occluded objects, they can improve the accuracy and reliability of object detection in complex environments, and better describe object pose and shape. Usually approaches that rely on monocular cameras are also less expensive and can be easily deployed in urban areas, if compared to the more complex *LIDAR* methods [12,20,32,38]. However, 3D object detection from monocular cameras still poses several challenges and shows significant limitations. First, the lack of depth information in 2D images makes it difficult to accurately estimate the size and position of objects; second, environmental factors, such as occlusions, shadows, and weather conditions, can also affect their operational accuracy and reliability.

To overcome such issues, recent advances in autonomous driving solutions have shown promising results, among which the most noteworthy works are *M3DRPN* [5], *Kinematic* [6], *SMOKE* [25], *Monodle* [26], and *FOC3D* [36], to name a few. Additionally, a growing number of datasets [7,10,17,19,27–29,35] are being adopted to further improve the effectiveness of this technology. Notwithstanding these advancements, this field remains an active area of research, and further investigation is necessary to strengthen the performance of monocular camera-based 3D object detection models. Applying such models developed for autonomous driving to roadside cameras is possible, since these cameras usually provide a wider coverage area and greater robustness to occlusion, remain stable for extended periods of time, and are more suitable for event recognition. However, since the scenario is different from that of vehicle use, several questions on their generalizability are open. To improve performance in this context, novel datasets, such as *Ko-PER* [34], *Rope3D* [39], *BAAI-VANJEE* [13], *BoxCars* [33], and *DAIR-V2X* [41] have been proposed, but most of them are not public.

Motivated by the above limitations, in this paper, we designed a novel synthetic dataset, hereafter named as *MonoRoadCam*, with the twofold aim of: a) facilitating the adaptation of 3D object detection methods for use on roadside cameras; b) examine in this context the performance of existing methods borrowed from autonomous driving, in a consistent and unified setting. To generate such a dataset, we opted for the *CARLA simulation environment* [14], for its ability to provide complex data that mimics real-world scenarios. We also employed the widely adopted *KITTI format* [17] in order to guarantee standardization and reproducibility of the evaluation tests. Our contribution is threefold:

– We generated a synthetic dataset for monocular 3D object detection from roadside cameras using the CARLA simulator, compliant with the KITTI format. To provide a fair evaluation, we removed the overlap between the training and validation sets by excluding sequence frames;

- We verified the reproducibility of existing state-of-the-art monocular 3D object detection approaches, originally proposed for autonomous driving, on the roadside context, sharing our framework publicly.
- We conducted a comparative study between 3D object detection datasets from roadside and frontal cameras, observing that state-of-the-art solutions from the autonomous driving domain result in significant potential yet crucial limitations when applied to monocular roadside camera images.

The rest of this paper is organized as follows. Section 2 outlines the research methodology. Section 3 illustrates the obtained results. Finally, Sect. 4 concludes the paper, highlighting the prominent future research directions.

2 Research Methodology

In this section, we describe the reproducibility process, which involves (i) surveying the existing datasets and 3D object detection methods for monocular cameras; (ii) carrying out an analysis of the context, also by collecting the original source codes and adapting them to our unified framework; (iii) generating the MonoRoadCam dataset and adopt it for evaluating the reproduced models.

2.1 Problem Formulation

Given a set of training images $I = \{i_1, i_2, ...i_n\}$ and calibration information P of a monocular camera, where P represents the projection matrix, each image $i \in I$ is represented as a set of 2D projected points. Suppose that $B = \{b_1, b_2, ...b_m\}$ represents a set of bounding boxes for all objects in the image in 3D space, where each $b_i \in B$ included object type C, in addition to $T = (tx, ty, tz)$, $D = (dx, dy, dz)$, and $O = (\vartheta, \Phi, \varphi)$ which represent the centroid, dimension, and orientation of the object, respectively. The goal is to optimize the parameter θ in order to solve $f(i, P, \theta) = B \ \forall i \in I$. Usually, convolution neural networks are used to provide the map, and the optimization is run on θ.

2.2 Paper Collection

In order to gather existing 3D datasets for the study, a systematic search has been conducted about the recent publications in computer vision-related top-tier conferences and journals, such as CVPR, ECCV, ICCV, and IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). Additionally, we searched relevant repositories, such as the Waymo Open Dataset, the KITTI Vision Benchmark Suite, and the ApolloScape dataset. Our search keywords included 3D object detection, monocular 3D object detection, roadside 3D dataset.

Not only datasets providing 3D object information using RGB cameras have been taken into consideration, but also those that used other methods in addition to RGB. Additionally, we limited our search to datasets designed for traffic monitoring using monocular cameras. We excluded datasets that focused on other

Table 1. Comparison of existing publicly available datasets for autonomous driving (AD) and roadside object detection, including their year of release, database type, whether the data is real or simulated, range of data, RGB resolution, number of RGB images, number of 3D boxes, presence of rain/night data, and availability to the public. The last row represents the dataset we propose in this paper.

Dataset	Year	Type	Source	Range	Resolution	Images	3D Boxes	Rain/ Night	Public?
Kitti [17]	2013	AD	Real	70 m	1392 × 512	15K	80K	No/No	Yes
KoPER [34]	2014	Roadside	Real	–	656 × 494	–	–	No/No	Yes
Apollscape [19]	2018	AD	Real	420 m	3384 × 2710	144K	70K	No/Yes	Yes
BoxCars [33]	2018	Roadside	Real	–	128 × 128	116K	116K	No/No	Yes
nuScenes [7]	2019	AD	Real	75 m	1600 × 900	1.4M	1.4M	Yes/Yes	Yes
Argoverse [10]	2019	AD	Real	200 m	1920 × 1200	22K	993K	Yes/Yes	Yes
H3D [28]	2019	AD	Real	100 m	1920 × 1200	27.7K	1M	No/No	Yes
A*3D [29]	2020	AD	Real	100 m	2048 × 1536	39K	230K	Yes/Yes	Yes
Waymo Open [35]	2020	AD	Real	75 m	1920 × 1080	230K	12M	Yes/Yes	Yes
DAiRV2X [41]	2021	AD/Other	Real	200 m	1920 × 1080	71K	1.2M	–/Yes	No
BAAI-VANJEE [13]	2021	Roadside	Real	–	1920 × 1080	5K	74K	Yes/Yes	No
ONCE [27]	2021	AD	Real	200 m	1920 × 1080	7M	417K	Yes/Yes	Yes
Rope3D [39]	2022	Roadside	Real	200 m	1920 × 1200	50K	1.5M	Yes/Yes	No
Ours	**2023**	**Roadside**	**Simulated**	**150 m**	**1280 × 384**	**15K**	**39.345K**	**No/Yes**	**Yes**

tasks or used different data collection methods. After conducting the search and filtering process, we identified *13* relevant datasets that met our criteria for inclusion in our study. Among them, *8* datasets were related to autonomous driving focusing on the frontal view of the road, *4* datasets were based on roadside cameras, and *1* dataset focused on autonomous driving and infrastructure. Table 1 summarizes their general characteristics.

As a second step of our study, we surveyed papers proposing monocular 3D object detection methods. Despite their efficiency, we excluded models that rely on LiDAR and point cloud data [12,20,32,38] from our study, given our focus on contexts with only monocular cameras. Additionally, we excluded models that heavily rely on external sub-networks for performing depth estimation [37] or pseudo point cloud generation [11], given the need of efficiency. Similarly to the dataset selection process, we targeted works from top-tier conferences and journals that propose an approach for monocular 3D object detection and that make that approach reproducible by sharing the source code. Based on these criteria, we were able to select four models: M3DPRN, Kinematic, SMOKE, and Monodel. All of them are autonomous driving-based models. During our search,

we also found two roadside models leveraging monocular cameras [43] and [9], but unfortunately the authors did not provide any source code.

2.3 Research Context Analysis

Based on datasets and the monocular 3D object detection models we have examined, we observed that approximately about 68% of them focused on autonomous driving, assuming that DAiRV2X is an autonomous driving dataset. However we also found that only 32% of the work focused on roadside cameras. These models can be useful in various applications, such as sports analysis, traffic monitoring, security systems, road safety, and wildlife monitoring. In addition to this, we found that this area lacks publicly available datasets until the date of this study. Rope3D [39], DAiRV2X [41], BAAI-VANJEE [13] are not publicly available. On the other side, we analyzed the training and testing datasets used by the state-of-the-art models surveyed in our study. From Table 2, we observed that most of surveyed 3D monocular object detection models still rely on the KITTI dataset for their training and testing, despite the availability of diverse publicly available datasets for autonomous driving, especially those that focus on the frontal view. This might be attributed to either the pioneering role of KITTI.

Table 2. Overview of the considered 3D object detection methods.

Method	Year	Type[1]	Status[2]	Datasets	Datasets size[3]
M3DRPN [5]	2019	AD	R	KITTI [17]	3,712-3,769-7,518
Kinematic [6]	2020	AD	R	KITTI [17]	3,712-3,769-7,518
SMOKE [25]	2020	AD	R	KITTI [17]	3,712-3,769-7,518
Monodle [26]	2021	AD	R	KITTI [17]	3,712-3,769-7,518
FCOS3D [36]	2021	AD	R	nuScenes [36]	700-150-150*
UrbanNet [9]	2021	Roadside	$\overline{R}$	Synthetic Only [9]	500-0-100
Zou et al. [43]	2022	Roadside	$\overline{R}$	Real + Synthetic [43]	Synthetic: 64,000/Real: 8,000

Type[1]: AD - Autonomous driving model; Roadside - Roadside model
Status[2]: R - Reproducible model; $\overline{R}$ - Non-Reproducible.
Datasets size[3]: Training set - Validation set - Testing set.
* These values represent the ratio of the scenes instead of the dataset size.

Given all this information, and due to the lack of publicly available datasets for roadside 3D object detection and the convenience of using the same format as in KITTI, we opted to generate our own synthetic dataset with a focus on roadside 3D object contexts and to format it as the KITTI dataset. This allowed us to evaluate state-of-the-art methods (M3DRPN, Monodle, SMOKE, Kinematic - see Table 2) with our dataset smoothly. We chose synthetic data as a reference since it is cost-effective in a preliminary phase and selected Carla [14] as our platform since it is built on a foundation for learning reinforcement and imitation models, making it as simple as possible to resemble the real world.

2.4 Methods Reproduction

Data Generation. We generated the synthetic dataset using the Carla 0.9.13 simulator and placed an RGB camera at an intersection area in TOWN5. Detailed information about the software and hardware specifications can be found in the source code repository (Camera type: RGB Camera; Image resolution: 1280×384; Camera location: x = 10 y = 0 z = 10; Camera pitch, yaw, roll: 0; Field of view: 120). Specifically, our synthetic dataset, MonoRoadCam, is composed only of car objects at the intersection area, and includes 7,481 development images and 7,518 test images with annotations provided in the same format as the KITTI dataset. For every frame of both the development and testing sets, we ensured that at least one object is present. Each object in the dataset is defined by its type, size, location, and orientation. For simplicity, we set the occlusion and truncation levels to 0. Notably, the development images are not sequential, and the test images only include 10 continuous frames. This diversity provides a rich set of training and testing data. We extended the diversity of the dataset by incorporating three weather conditions: night, cloudy, and sunny. The statistics for the development and test sets are summarized in Table 3.

Table 3. Object size statistics in our dataset and in KITTI (car objects only).

Dataset	Statistic	Height [m]	Width [m]	Length [m]
KITTI (car object)	Average	1.53	1.63	3.53
Ours (car object)	Average	1.73	1.86	4.47
	Std. Dev.	0.49	0.50	1.38
	Min	1.20	0.33	1.49
	Max	3.83	2.89	8.47

Data Pre-processing. We generated the 3D boxes automatically in Carla. However to ensure that the data is free from any imprecise boxes, we performed data preprocessing in three steps. Firstly, we removed the boxes placed outside the road by determining the boundary of the road. Secondly, we removed too small boxes. Finally, we replaced images that did not contain any objects with other images. Figure 1 explains the data preprocessing phase in detail.

Model Creation. In our experiment, we evaluated two types of monocular 3D object detection methods: anchor-based and keypoint-based.

The anchor-based methods, Kinematic [6] and M3DRPN [5], aim to improve the accuracy of 3D estimation from a monocular camera. Kinematic incorporates uncertainty reduction, Kalman filtering, and ego-motion to extract scene dynamics. M3DRPN is built on the Faster R-CNN [31] concept and uses depth-aware concepts to improve accuracy. Both M3DRPN and Kinematic use predetermined 3D bounding boxes (i.e., *anchors*) and estimate the deviation from the anchor using offsets. On the other hand, the keypoint-based methods, Monodel [26] and SMOKE [25], do not rely on predetermined bounding boxes. Monodel focuses on

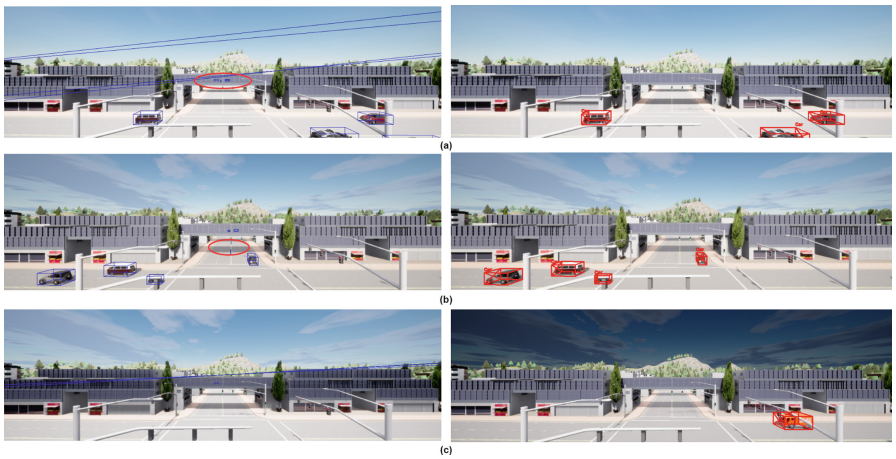

Fig. 1. Preprocessing steps for the images included in our dataset. In (a), we removed all the boxes outside the road. In (b), we removed small boxes by setting a threshold. In (c), we detected images not including any objects and replaced them. Blue (red) bounding boxes denote images before (after) preprocessing. (Color figure online)

improving dimension estimation and considering localization errors as a source of 3D detection inaccuracies. SMOKE estimates 3D objects directly without estimating the 2D bounding box. Both methods are anchor-free and use DLA34 [40] as a backbone. Despite the promising results, we decided not to include FCOS3D [36], which separates the 2D and 3D attributes of objects and redefines the centerness of objects based on the 3D center, since it required substantial steps to receive data in our required format, going beyond the scope of this study.

In addition to the monocular 3D object detection methods, we also examined two roadside methods. The first method [43] proposed 3D object detection and tracking using the point detection concept, then estimating the object's 3D pose and size. It predicts the object's 3D bottom center and uses a pre-calibrated plane-to-plane homography to lift the prediction to 3D space. The second method, UrbanNet [9], incorporated urban maps into the image to provide additional information to improve 3D estimation. Both these methods used synthetic datasets. Specifically, the first method used CARLA and UrbanNet used Grand Theft Auto V and KITTI. We excluded these methods from our analysis due to the lack of public source code and data set.

Evaluation. In order to reproduce monocular models in both the KITTI and MonoRoadCam datasets, we followed the same training/validation split protocol as proposed in [11], which is widely accepted in the field as a benchmark for evaluating the performance of monocular models. This protocol consists of 3,712 training and 3,769 validation images, and is commonly used to evaluate the performance of monocular models on the KITTI dataset. For our MonoRoadCam dataset, we confirmed that there were no sequential frames, but we still applied the same split protocol as in KITTI to unify the numbers of training and validation sets. This allowed us to perform a fair comparison.

For each trained model, we computed the evaluation metrics proposed and used for the KITTI dataset. Specifically, we included the 11-point and 40-point recall interpolated Average Precision (AP_{11} and AP_{40}) and the Average Orientation Similarity (AOS), which is used to measure the detector performance on rotated rectangle detection. These metrics are defined as follows.

$$AP_{11} = \frac{1}{11} \sum_{r \in 0, 0.1, \ldots, 1} p_{interp}(r), \tag{1}$$

where $p_{interp}(r)$ is the maximum precision for any recall value $r' \geq r$.

$$AP_{40} = \frac{1}{40} \sum_{r \in 0, 0.025, \ldots, 1} p_{interp}(r), \tag{2}$$

where $p_{interp}(r)$ is the maximum precision for any recall value $r' \geq r$.

$$AOS = \frac{1}{11} \sum_{r \in 0, 0.1, \ldots, 1} max_{r' \geq r} s(r') \tag{3}$$

where $s(r)$ is the orientation similarity.

$$s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \delta_i \cdot \frac{1}{2}(cos \Delta\theta_i^{(i)} + 1) \tag{4}$$

where $D(r)$ is all object detection at recall rate r, δ_i is a binary variable that is set to 1 if detection i has been assigned to a ground truth bounding box (overlaps by at least 50%), and $\Delta\theta_i^{(i)}$ is the difference in angle between the estimated and ground truth orientation of detection.

It should be noted that AP_{40} provides a more fine-grained evaluation than AP_{11}, by computing the precision at 40 different recall levels. AP_{11} computes precision at only 11 recall levels. Therefore, we used AP_{40} to perform a more detailed comparison of the different models. We used AP_{11} only for M3DRPN.

3 Experimental Results

3.1 RQ1: Status of Reproducibility

In Table 4, we report the full reproducibility results under the KITTI validation set for Kinematic, Monodle, M3DRPN, and SMOKE, together with the original results reported in their respective papers on the same validation set. Regarding M3DRPN, we conducted the same experiment as reported in the paper, using AP with 11 recall points, whereas for the other methods, we used 40 recall points for the AP calculation for the car object.

Our reproduced results showed a drop in performance for all the evaluated methods when compared to the same results reported in the original papers. However, on average, the decrease was not substantial, except for the SMOKE method. We conjecture that the decrease for the latter can be attributed to a misalignment in hyper-parameter values during training and decided to carefully consider this aspect while training models with SMOKE on our dataset.

Table 4. Comparison between the reproduced (*Ours*) and the original (*Orig*) results for the Kinematic, Monodle, M3DRPN, and SMOKE methods on the KITTI dataset, within the car object detection task. We report Average Precision (AP) values for both 3D and Bird's Eye (BEV) views, both at easy, moderate, and hard levels. (*) denotes AP for 11 recall points (40 recall points are used where not otherwise specified).

Method	AP_{3D} ($IoU \geq 0.7$)									AP_{BEV} ($IoU \geq 0.7$)								
	Easy			Mod			Hard			Easy			Mod			Hard		
	Ours	Orig	Gap %	Ours	Orig	Gap %	Ours	Orig	Gap %	Ours	Orig	Gap %	Ours	Orig	Gap %	Ours	Orig	Gap %
M3DRPN* [5]	14.37	20.40	−29.56	11.67	16.48	−29.19	9.23	13.34	−30.81	20.94	26.36	−20.56	15.35	21.15	−27.42	16.72	17.14	−2.45
Kinematic [6]	17.76	19.46	−8.74	13.45	14.10	−4.61	10.65	10.45	1.91	25.41	27.83	−8.70	18.79	19.72	−4.72	15.16	15.10	−0.40
SMOKE [25]	0.59	14.76	−96.00	0.58	12.85	−95.49	0.36	11.50	−96.87	1.60	19.99	−92.00	1.26	15.61	−91.93	1.21	15.28	−92.08
Monodle [26]	11.40	17.45	−34.67	9.10	13.66	−33.38	7.55	11.68	−35.36	16.97	24.97	−32.04	13.26	19.33	−31.40	11.89	17.01	−30.10

3.2 RQ2: Influence of the Context

The context in which 3D object detection methods are applied can have a significant impact on their performance. Therefore, we examined the performance of the methods, training models from scratch on our synthetic dataset. We used the Average Precision (AP) metric for both 3D object detection and bird's eye view (BEV) object detection. The results were computed separately for easy, moderate, and hard difficulty levels to provide a comprehensive evaluation.

Table 5. Comparison between the results obtained on the autonomous driving (AD) scenario (i.e., *Ours* in Table 4, calculated on the KITTI dataset) and on the roadside cameras (RC) scenario (i.e., our synthetic dataset) by the considered models, within the car object detection task. We report the Average Precision (AP) for both 3D and Bird's Eye (BEV) views, at easy, moderate, and hard levels. (*) denotes AP for 11 recall points (40 recall points are used where not otherwise specified, as in Table 4).

Method	AP_{3D} ($IoU \geq 0.7$)									AP_{BEV} ($IoU \geq 0.7$)								
	Easy			Mod			Hard			Easy			Mod			Hard		
	AD	RC	Gap %	AD	RC	Gap %	AD	RC	Gap %	AD	RC	Gap %	AD	RC	Gap %	AD	RC	Gap %
M3DRPN* [5]	14.37	51.14	255.90	11.67	50.43	332.13	9.23	50.43	446.40	20.94	54.12	158.45	15.35	53.73	250.03	16.72	53.73	221.35
Kinematic [6]	17.76	56.49	218.07	13.45	54.15	302.60	10.65	54.15	408.45	25.41	59.40	133.77	18.79	57.27	204.80	15.16	57.27	277.77
SMOKE [25]	0.59	0.15	−74.58	0.58	1.30	124.14	0.36	1.30	261.10	1.60	0.61	−61.88	1.26	6.20	392.10	1.21	6.20	412.40
Monodle [26]	11.40	10.78	−5.44	9.10	9.91	8.90	7.55	9.91	31.26	16.97	11.89	−29.94	13.26	12.59	−5.05	11.89	12.59	5.89

Table 6. Results on the validation set of our synthetic dataset using the Kinematic, Monodle, M3DRPN, and SMOKE methods. The table reports Average Precision (AP) with 40 recall points in 3D and Bird's Eye View (BEV) - the same results reported in Table 5 - column *RC*, but organized here to ease the comparison across models - and the Average Orientation Similarity (AOS) for the easy, moderate, and hard levels.

Method	AP (IoU>=0.7)						AOS		
	Easy		Mod		Hard		Easy	Mod	Hard
	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}			
M3DRPN [5]	51.14	54.12	50.43	53.73	50.43	53.73	46.01	46.09	46.09
Kinematic [6]	**56.49**	**59.40**	**54.15**	**57.27**	**54.15**	**57.27**	45.52	46.73	46.73
SMOKE [25]	0.15	0.61	1.30	6.20	1.30	6.20	2.25	8.56	8.56
Monodle [26]	10.78	11.89	9.91	12.59	9.91	12.59	**61.53**	**63.75**	**63.75**

Fig. 2. Qualitative comparison across models on the validation set of our dataset.

Comparing results across contexts (autonomous driving and roadside views) in Table 5, it can be observed that most of the methods perform better on our synthetic dataset (RC) compared to our experiment in the original KITTI dataset (AD). For example, the Easy mode in the Kinematic method exhibited a substantial gap, with a score of 56.49 in our dataset compared to 17.76 in KITTI. However, it is crucial to consider the variations in object diversity between the two datasets, particularly concerning object boundaries in our CARLA environment. We specifically tested the intersection area of a single scene. Although we took precautions to prevent overlap between the training and validation sets, and even in the training and validation phases, by excluding sequential frames, it is possible that the inherent boundary characteristics from our CARLA setup could still influence the results. Furthermore, it can be observed that the performance of all methods decreases as the difficulty level increases in both datasets. This is expected as the difficulty levels correspond to objects with smaller sizes (we avoided occlusion and truncation in our dataset).

Comparing results across models (Kinematic, Monodle, M3DRPN, SMOKE) in Table 6, it can be observed that Kinematic achieved the highest performance in both AP 3D and AP BEV, with the best result being 59.40 in the Bird's Eye View easy mode. M3DRPN ranked second with strong AP scores. Conversely, despite having lower AP scores compared to Kinematic and M3DRPN, Monodle showcased impressive results in terms of AOS (63.75) under moderate and hard difficulty levels. SMOKE reported the lowest overall performance, consistently scoring lower in all metrics and difficulty levels among the four methods. Based on such results, we concluded that the models exhibit effectiveness in the roadside scenario of our synthetic dataset, especially when Kinematic is used.

3.3 RQ3: Qualitative Inspection

We finally conducted a qualitative comparison of the models, by employing specific challenging images chosen from the validation set. Images in Fig. 2 cover various scenarios involving big and small cars as well as cars in close proximity.

We found that Kinematic exhibited good projection accuracy, by accurately localizing objects within the scenes. On the other side, M3DRPN displayed a few false positives in some images (see columns 1–3). When it comes to Monodle, we noticed limitations in terms of IOU scores in columns 1–2, indicating that it struggles to precisely capture object boundaries. Furthermore, Monodle generated a false negative in column 3. Interestingly, both M3DRPN and Monodle detected truncated objects in column 2. As for SMOKE, we observed some limitations in terms of IOU scores, false positives, and false negatives. Notably, the latter faced noticeable challenges in accurately detecting larger cars.

4 Conclusions and Future Work

In this study, we shed a light on the scarcity of publicly available datasets for 3D object detection from monocular cameras in roadside contexts. To address this issue, we introduced a synthetic dataset generated through the CARLA simulator, which is compatible with the popular KITTI format and can be seamlessly integrated into existing frameworks. Furthermore, we showed the feasibility of our dataset by verifying the reproducibility of state-of-the-art monocular autonomous driving models on roadside contexts, yielding promising initial results.

Our findings suggest that our synthetic dataset could serve as a valuable resource for researchers and practitioners in the field of autonomous driving, facilitating the development and evaluation of 3D object detection algorithms for roadside scenarios. Therefore, as next steps, from a data perspective, we plan to extend the generated dataset with more examples and situations and to explore innovative ways for gathering real-world annotated datasets. From a methodological perspective, we plan to devise models that can lead to more effective and efficient computation under the considered roadside scenario. Finally, to assess the impact of our work on the real world, we plan to run applicative studies involving local municipalities.

Acknowledgements. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union - NextGenerationEU. Project Code ECS0000038 - Project Title e.INS Ecosystem of Innovation for Next Generation Sardinia - CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the MUR.

References

1. Atzori, A., Barra, S., Carta, S., Fenu, G., Podda, A.S.: HEIMDALL: an AI-based infrastructure for traffic monitoring and anomalies detection. In: 19th IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2021, Kassel, Germany, 22–26 March 2021, pp. 154–159. IEEE (2021). https://doi.org/10.1109/PerComWorkshops51409.2021.9431052

2. Atzori, A., Fenu, G., Marras, M.: Explaining bias in deep face recognition via image characteristics. In: IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, 10–13 October 2022, pp. 1–10. IEEE (2022). https://doi.org/10.1109/IJCB54206.2022.10007937
3. Atzori, A., Fenu, G., Marras, M.: Demographic bias in low-resolution deep face recognition in the wild. IEEE J. Sel. Top. Signal Process. **17**(3), 599–611 (2023). https://doi.org/10.1109/JSTSP.2023.3249485
4. Balia, R., Barra, S., Carta, S., Fenu, G., Podda, A.S., Sansoni, N.: A deep learning solution for integrated traffic control through automatic license plate recognition. In: Gervasi, O., et al. (eds.) ICCSA 2021, Part III. LNCS, vol. 12951, pp. 211–226. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86970-0_16
5. Brazil, G., Liu, X.: M3D-RPN: monocular 3D region proposal network for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), 27 October–2 November 2019, pp. 9286–9295. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00938
6. Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3D object detection in monocular video. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part XXIII. LNCS, vol. 12368, pp. 135–152. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_9
7. Caesar, H., et al.: nuScenes: a multimodal dataset for autonomous driving. CoRR abs/1903.11027 (2019). https://arxiv.org/abs/1903.11027
8. Cao, J., Cholakkal, H., Anwer, R.M., Khan, F.S., Pang, Y., Shao, L.: D2Det: towards high quality object detection and instance segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 11482–11491. Computer Vision Foundation/IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01150
9. Carrillo, J., Waslander, S.L.: UrbanNet: leveraging urban maps for long range 3D object detection. In: 24th IEEE International Intelligent Transportation Systems Conference, ITSC 2021, Indianapolis, IN, USA, 19–22 September 2021, pp. 3799–3806. IEEE (2021). https://doi.org/10.1109/ITSC48978.2021.9564840
10. Chang, M., et al.: Argoverse: 3D tracking and forecasting with rich maps. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 8748–8757. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.00895
11. Chen, X., et al.: 3D object proposals for accurate object class detection. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, 7–12 December 2015, Montreal, Quebec, Canada, pp. 424–432 (2015). https://proceedings.neurips.cc/paper/2015/hash/6da37dd3139aa4d9aa55b8d237ec5d4a-Abstract.html
12. Chen, Y., Liu, S., Shen, X., Jia, J.: Fast point R-CNN. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), 27 October–2 November 2019, pp. 9774–9783. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00987
13. Deng, Y., et al.: BAAI-VANJEE roadside dataset: towards the connected automated vehicle highway technologies in challenging environments of china. CoRR abs/2105.14370 (2021). https://arxiv.org/abs/2105.14370

14. Dosovitskiy, A., Ros, G., Codevilla, F., López, A.M., Koltun, V.: CARLA: an open urban driving simulator. In: 1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, 13–15 November 2017, Proceedings. Proceedings of Machine Learning Research, vol. 78, pp. 1–16. PMLR (2017). https://proceedings.mlr.press/v78/dosovitskiy17a.html
15. Fenu, G., Marras, M.: Controlling user access to cloud-connected mobile applications by means of biometrics. IEEE Cloud Comput. **5**(4), 47–57 (2018). https://doi.org/10.1109/MCC.2018.043221014
16. Fenu, G., Marras, M., Medda, G., Meloni, G.: Causal reasoning for algorithmic fairness in voice controlled cyber-physical systems. Pattern Recognit. Lett. **168**, 131–137 (2023). https://doi.org/10.1016/j.patrec.2023.03.014
17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012, pp. 3354–3361. IEEE Computer Society (2012). https://doi.org/10.1109/CVPR.2012.6248074
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR abs/1703.06870 (2017). https://arxiv.org/abs/1703.06870
19. Huang, X., et al.: The ApolloScape dataset for autonomous driving. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 954–960. Computer Vision Foundation/IEEE Computer Society (2018). https://doi.org/10.1109/CVPRW.2018.00141
20. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: PointPillars: fast encoders for object detection from point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 12697–12705. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.01298
21. Law, H., Deng, J.: CornerNet: Detecting Objects as Paired Keypoints. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part XIV. LNCS, vol. 11218, pp. 765–781. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_45
22. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), 27 October–2 November 2019, pp. 6053–6062. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00615
23. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 2999–3007. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.324
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part I. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
25. Liu, Z., Wu, Z., Tóth, R.: SMOKE: single-stage monocular 3D object detection via keypoint estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, 14–19 June 2020, pp. 4289–4298. Computer Vision Foundation/IEEE (2020). https://doi.org/10.1109/CVPRW50498.2020.00506

26. Ma, X., et al.: Delving into localization errors for monocular 3D object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, 19–25 June 2021, pp. 4721–4730. Computer Vision Foundation/IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00469
27. Mao, J., et al.: One million scenes for autonomous driving: ONCE dataset. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual (2021). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/67c6a1e7ce56d3d6fa748ab6d9af3fd7-Abstract-round1.html
28. Patil, A., Malla, S., Gang, H., Chen, Y.: The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes. In: International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, 20–24 May 2019, pp. 9552–9557. IEEE (2019). https://doi.org/10.1109/ICRA.2019.8793925
29. Pham, Q.H., et al.: A*3D dataset: towards autonomous driving in challenging environments. In: Proceedings of the International Conference in Robotics and Automation (ICRA) (2020)
30. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 779–788. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.91
31. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, 7–12 December 2015, Montreal, Quebec, Canada, pp. 91–99 (2015). https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html
32. Shi, S., et al.: PV-RCNN: point-voxel feature set abstraction for 3D object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 10526–10535. Computer Vision Foundation/IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01054
33. Sochor, J., Špaňhel, J., Herout, A.: BoxCars: improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance. IEEE Trans. Intell. Transp. Syst. **PP**(99), 1–12 (2018). https://doi.org/10.1109/TITS.2018.2799228
34. Strigel, E., Meissner, D.A., Seeliger, F., Wilking, B., Dietmayer, K.: The Ko-PER intersection laserscanner and video dataset. In: 17th International IEEE Conference on Intelligent Transportation Systems, ITSC 2014, Qingdao, China, 8–11 October 2014, pp. 1900–1901. IEEE (2014). https://doi.org/10.1109/ITSC.2014.6957976
35. Sun, P., et al.: Scalability in perception for autonomous driving: waymo open dataset. CoRR abs/1912.04838 (2019). https://arxiv.org/abs/1912.04838
36. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: fully convolutional one-stage monocular 3D object detection. In: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, 11–17 October 2021, pp. 913–922. IEEE (2021). https://doi.org/10.1109/ICCVW54120.2021.00107

37. Xu, B., Chen, Z.: Multi-level fusion based 3D object detection from monocular images. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 2345–2353. Computer Vision Foundation/IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00249
38. Yan, Y., Mao, Y., Li, B.: SECOND: sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018). https://doi.org/10.3390/s18103337
39. Ye, X., et al.: Rope3D: the roadside perception dataset for autonomous driving and monocular 3D object detection task. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022, pp. 21309–21318. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.02065
40. Yu, F., Wang, D., Darrell, T.: Deep layer aggregation. CoRR abs/1707.06484 (2017). https://arxiv.org/abs/1707.06484
41. Yu, H., et al.: DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection. CoRR abs/2204.05575 (2022). https://doi.org/10.48550/arXiv.2204.05575
42. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. CoRR abs/1904.07850 (2019). https://arxiv.org/abs/1904.07850
43. Zou, Z., et al.: Real-time full-stack traffic scene perception for autonomous driving with roadside cameras. In: 2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, 23–27 May 2022, pp. 890–896. IEEE (2022). https://doi.org/10.1109/ICRA46639.2022.9812137