Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task

Xiaoqing Ye\textsuperscript{1,*} Mao Shu\textsuperscript{1,*} Hanyu Li\textsuperscript{1} Yifeng Shi\textsuperscript{1} Yingying Li\textsuperscript{1} Guangjie Wang\textsuperscript{2} Xiao Tan\textsuperscript{1,†} Errui Ding\textsuperscript{1}

\textsuperscript{1}Baidu Inc. \hspace{1em} \textsuperscript{2}China University of Mining and Technology

Abstract

Concurrent perception datasets for autonomous driving are mainly limited to frontal view with sensors mounted on the vehicle. None of them is designed for the overlooked roadside perception tasks. On the other hand, the data captured from roadside cameras have strengths over frontal-view data, which is believed to facilitate a safer and more intelligent autonomous driving system. To accelerate the progress of roadside perception, we present the first high-diversity challenging Roadside Perception 3D dataset—Rope3D from a novel view. The dataset consists of 50k images and over 1.5M 3D objects in various scenes, which are captured under different settings including various cameras with ambiguous mounting positions, camera specifications, viewpoints, and different environmental conditions. We conduct strict 2D-3D joint annotation and comprehensive data analysis, as well as set up a new 3D roadside perception benchmark with metrics and evaluation devkit. Furthermore, we tailor the existing frontal-view monocular 3D object detection approaches and propose to leverage the geometry constraint to solve the inherent ambiguities caused by various sensors, viewpoints. Our dataset is available on https://thudair.baai.ac.cn/rope.

1. Introduction

Autonomous driving plays a crucial role in helping reduce traffic accidents and improve transportation efficiency. Current perceptual systems mainly equip the moving vehicle with LiDAR or camera sensors. Owing to the movement, the vehicle perceptual system can not observe surroundings for a long period. In addition, since the mounted sensor is relatively low (usually on the top of a vehicle), the perceptual range is comparatively limited and is vulnerable to occlusion. On the contrary, the data captured from roadside cameras has its inherent strengths in terms of robustness to occlusion and long-time event prediction, since they are collected from cameras mounted on poles a few meters above the ground. The comparisons between two different views of data are depicted in Fig. 1.

The importance of roadside perception is listed as follows: (1) Cooperative to Autonomous driving (AD). AD still faces safety challenges and uncontrolled threats due to blind spots. Instead, the roadside view can cover the blind spots for two extra advantages over car views: a long-range global perspective to extend vehicles’ perception field spatially and temporally and global trajectory prediction for...
safety. For example, a pedestrian walking behind a parked vehicle might suddenly crash into a moving vehicle since vehicle sensors fail to detect abrupt changes in the environment owing to the limited perceptual range or heavy occlusion. On the contrary, the roadside view is capable of behavior prediction timely. (2) Global perception. Further objects are occluded (even with 360° sensors) by closer objects in existing car-view datasets, causing blind spots. Thanks to roadside cameras mounted overhead, the invisible region is now visible. Besides, Autonomous vehicles (AV) can be informed to choose a faster lane when having a dead car in the queue since the roadside view perceives globally. (3) Cost-efficient. In terms of cost, it is worthy for ensuring safety by cooperative perception and cost-efficient since information from roadside cameras can broadcast to all surrounding AVs. (4) Intelligent traffic control. The roadside perception also facilitates smart traffic control and flow management. The critical contribution of roadside perceptual systems in facilitating a safer and more intelligent autonomous driving system has been acknowledged in many works [8, 34, 41].

However, existing researches on the roadside perceptual ability focus only on 2D tasks such as 2D detection and tracking, the ability of 3D localization is still under-explored [29, 30, 39]. In this work, we focus on monocular 3D detection that localizes objects in 3D space from a single image. Although abundant perception datasets have been published to fuel the development in autonomous driving from vehicle view, such as KITTI [12], nuScenes [5], A*3D [32] and Waymo [38], none of them is designed particularly for the overlooked roadside 3D perception task. We hence release the first large-scale high-diversity roadside Perception Dataset (Rope3D), with the hope of bridging this gap. Compared with the existing vehicle view datasets, the roadside perceptual data can be different in three ways. First, the ambiguity lies everywhere due to various cameras specifications such as distinct pitch angles of the viewpoint, mounting heights as well as various roadside environments, which increases the difficulty of monocular 3D detection tasks to a great extent. Second, since the roadside cameras are mounted on the poles instead of on top of the vehicle in frontal view, thus the assumption of the camera’s optical axis being parallel to the ground is no more valid, leading to the incompatibility of directly applying the existing monocular 3D detection approaches using this prior. Third, due to a much larger sensible range of the roadside perceptual system, a larger number of objects are expected to observe in roadside view, increasing the density and difficulty of a perceptual system. All these differences prevent directly applying most existing 3D detection methods. We hence tailor existing monocular 3D object detection methods to the roadside application.

To summarize, our contributions are as follows:

• We present the first challenging high-diversity roadside dataset termed “Rope3D”, consisting of 50k images and over 1.5M 3D objects collected across a variety of lighting conditions (daytime / night / dusk), different weather conditions (rainy / sunny / cloudy), and distinct road scenes with different camera specifications like focal length and viewpoints.

• We specially tailor current frontal-view monocular 3D detection methods to deal with the roadside view data and conduct a comprehensive study with the new 3D detection metrics particularly designed for roadside 3D detection tasks, hoping to facilitate the development of monocular 3D perception tasks in roadside scenarios.

2. Related work

2.1. Frontal-view Autonomous Driving Datasets

Recent years have witnessed great progress in autonomous driving, thanks to a great number of large-scale traffic scene datasets. [2, 10, 11, 17, 26, 32, 44] As a pioneer work, KITTI [12] provides multimodal data and opens a leader board for a variety of tasks. Although the raw point clouds from LiDAR is 360° around the collecting vehicle, the annotation is only conducted within the overlapping frontal view of the camera. To tackle the limitation, the Honda Research Institute 3D Dataset (H3D) [31] supplies a total of 1.1M 3D boxes in full 360 view, which accelerates full-surround multi-object detection and tracking. Another 360° view multimodal 3D detection dataset is introduced in nuScenes [5], providing over 1.4M annotated 3D boxes in 1000 scenes, including nighttime and rainy conditions. The ApolloScape and ApolloCar3D [16, 37] focus on the pixel-wise semantic segmentation task, including scene parsing, 3D car instance, lane segmentation tasks. The Argoverse dataset [7] designs for vehicle perception tasks such as 3D tracking and motion forecasting. The Waymo Open Dataset [38] consists of over 1000 scenes and 12M 3D boxes in urban and suburban scenarios, under various weather and lighting conditions. Although the above-mentioned datasets have fueled the innovation on autonomous driving, they are all designed for vehicle view perception. However, there is a lack of a dataset helping us to effectively carry out research on 3D localization under roadside surveillance cameras. A related work is from BoxCars [36], which performs fine-grained vehicle recognition by estimating the projected vertices of 3D bounding box on the image rather than real-world location, size, and orientation. Another contemporary work [49] focuses on the LiDAR-based 3D detection task. The comparisons between our roadside 3D dataset and the previous AD datasets are listed in Table 1.

2.2. Monocular 3D Object Detection

Though challenging, monocular-based 3D detection is still an attractive solution especially in autonomous driving
systems, where the method predicts the 3D bounding boxes from a single image [1,9,18–20,22,24,28,45]. Monocular 3D detection methods can be divided into three categories. 

(1) Anchor-based. Methods in this category exploit a series of predefined 3D bounding box with a location called “anchor” and estimate the offset w.r.t the anchor. M3D-RPN [3] leverages a 3D region proposal network and the geometric constrains of 2D and 3D perspectives to directly regress the 3D location and size. Kinematic3D [4] further extends M3D-RPN by leveraging 3D kinematics from monocular videos to improve the overall localization. 

(2) Keypoint-based. Many attempts [23,25,33,40,50] have been made to directly regress the keypoints, and then estimate 3D bounding box size and location from the image position of keypoints by optimization e.g., RTM-3D [21] and MonoGRNet [33]. 

(3) Pseudo-Lidar / depth based. Extra depth estimation modules and/or point cloud guidance are employed to alleviate the lack of accurate depth information. The pioneering work pseudo-LiDAR [42,43,47] imitates the LiDAR-based methods by utilizing off-the-shelf depth estimators to convert image pixels into pseudo-LiDAR point clouds, and employs LiDAR-based approaches for further detection. DA-3Ddet [46] adapts the features from unsound image-based pseudo-LiDAR domain to reliable LiDAR domain for guidance to boost the monocular performance. UrbanNet [6] utilizes the urban 3D map, including driving lanes, elevation, and slope as prior for assisting the 3D detection task. The existing monocular 3D detection methods are mainly designed for processing vehicle view data. Due to the domain gap and the distribution shift, a question raise naturally is whether these methods still applicable for roadside applications and if not how can we tailor these methods to the new scenario.

3. The Roadside Perception Dataset

3.1. Specifications

Sensors Setup. The roadside data collection is conducted by two sets of sensors, one is the roadside cameras mounted
on the pole or beside the traffic light; the other is the LiDAR equipped on a parked/driving vehicle to obtain the 3D point clouds of the same scene. For sensor synchronization, we adopt the nearest time matching strategy to find the pairs of image and point clouds within 5 milliseconds error.

- Roadside cameras: RGB with 1920 × 1080 resolution, 30-60Hz capture frequency and 1/1.8” CMOS sensor.
- LiDAR: (1) HESAI Pandar 40P, 40 laser beams, 10/20Hz capture frequency, ±2cm accuracy, 360° horizontal FOV, -25°~+15° vertical FOV, ≤ 200m range. (2) Jaguar Prime from Innovusion: 300 beams, 6-20 FPS with ±3cm accuracy, 100° horizontal FOV, 40° vertical FOV, ≤ 280m range.

Coordinate Systems and Calibration. There are three coordinate systems used in the dataset: the World Coordinate (i.e., the Universal Transverse Mercator coordinate system (UTM Coord.), the Camera Coordinate, as well as the LiDAR Coordinate.

To obtain the reliable ground truth 2D-3D joint annotation, the calibration between different sensors is required. First, the camera is calibrated to obtain the intrinsics by detecting the chessboard patterns. Then the Lidar-to-World calibration is conducted by the vehicle localization module to obtain the high definition (HD) map in UTM Coord. For World-to-Camera calibration, we first project the HD map which contains lane and crosswalk endpoints onto the 2D image to obtain the raw transformation. A bundle adjustment refinement is followed to derive the final transformation. Then Lidar-to-Camera transformation can be obtained by simply multiplying Lidar-to-World and World-to-Camera transformations.

After obtaining the transformation between the three coordinate systems, we can easily compute the ground equation \( G(\alpha, \beta, \gamma, d) \) by fitting the ground points \([x, y, z] \) to the ground plane in the camera coordinate, with \( \alpha x + \beta y + \gamma z + d = 0 \).

3.2. Data Collection and Annotation

After obtaining the intrinsics as well as the LiDAR-to-Camera transformation, we can collect the 2D-3D data. We choose various roadside cameras and let a LiDAR-equipped vehicle park or drive around. To keep the high diversity and complexity of the real environment, we collect more than 50k image frames at different times (daytime, night, dawn/dusk), different weather conditions (sunny, cloudy, rainy), different densities (crowded, normal, less traffic), different distributions of traffic elements and so on. There are totally 13 object classes with their corresponding category, 2D properties (occlusion, truncation) and the 7-DOF 3D bounding box: Location \((x, y, z)\). Size (width-\(W\), length-\(L\), height-\(H\)), Orientation (the heading angle-\(\theta\)). The full pipeline is illustrated in Fig. 3. (1)First after obtaining the 3D point clouds and the 2D image (they are within the same space but differ in the viewpoint), we first annotate the 3D bounding boxes directly on the 3D point clouds. (2) Simultaneously, the annotated 3D bounding boxes will be projected on the 2D image plane, see the top part of Fig. 3(c). We adjust the 3D parameters so that the projected points align with the 2D instance and mainly cover it. (3) For 2D box annotations, if the objects are scanned by the laser, their 2D box labels in the image are the minimum bounding box of the amodal projections of the eight 3D corners. For objects that are heavily occluded or too far to be detected by laser, 2D complementary labeling is conducted to label 2D bounding boxes directly in the image and leave its 3D annotations empty, see the bottom part of Fig. 3(c).

3.3. Statistics and Analysis

The collected images have high diversity and inherent ambiguity due to different settings of pitch angles, height, and camera types as well as the manifold scenes. Two levels of categories are adopted in the dataset. The coarse-grained level mainly focuses on the most common traffic elements: Car, Big Vehicle, Pedestrian, and Cyclist. To be more fine-grained, Car includes car and van, Big Vehicle can be further divided into truck and bus, and meanwhile, Cyclist can be subdivided into cyclist, motorcyclist, barrow, and tricyclist, since they are driving non-motor vehicles. We have annotated 13 classes, in addition to the above-mentioned categories, there are four extra classes: ‘traffic cones’, ‘triangle_plate’, ‘unknown-unmovable’, ‘unknown-movable’. The following statistics are mainly on the coarse and fine-grained classes.

Quantity distribution. We first give an overview analysis of the dataset on the number of 2D and 3D objects in Fig. 4. As is stated, 2D objects are more than 3D objects since some objects are not scanned by the LiDAR laser, so they only have 2D annotations. We give the detailed number of coarse level and fine-grained levels of the categories, corresponding to (a) and (b) in Fig. 4.

Depth distribution. Besides, we analyze the depth distribution of coarse-categories in Fig. 5. The depth of the captured 3D objects can range from within 10m to over 140m. Most objects lie between 60 and 80 meters.

Density. The density can be a key factor that affects the capability of 3D perception. Thus, we analyze the density of the dataset from two aspects in Fig. 6. From the global level, we show the 2D and 3D annotated number of each image in (a). The samples can be up to more than one hundred. Compared to other datasets whose densities are KITTI [12]: 5.3, nuScenes [5]: 9.7 and A*3D [32]: 5.9, our dataset has a much higher density (34 and 24 for 2D/3D per image). From the view of coarse categories, the numbers of 3D samples per frame are shown in Fig. 6 (b). The ‘Car’ category is relatively evenly distributed in densities whereas there are less than 10 big vehicles in each image.

21344
Figure 3. The data acquisition and labeling pipeline. Our platform takes the captured roadside images and the point clouds scanned by the LiDAR mounted on a parked/driving vehicle as input. After calibration and alignment between various sensors, the transformations between LiDAR, the world, and the camera are obtained, as well as the ground plane equation and intrinsics. The 2D-3D joint annotation is carried out by projecting the point clouds onto the images and adjusting the 3D bounding boxes manually to fit the 2D instance. For objects that are not scanned by the laser, the 2D complementary labeling is performed on the images only. For example in (d), some objects only have white 2D bounding boxes and no 3D colored annotations due to a lack of 3D points.

Figure 4. The quantities of: (a) coarse level, (b) fine-grained level.

Figure 5. The depth distribution of coarse categories.

Occlusion and truncation Analysis. Next, we annotate three levels for occlusion and truncation attributes. For occlusion, Level 0 denotes no occlusion, 1 and 2 means less / more than 50% occlusion. For truncation attribute, Level 0 means no truncation, 1 and 2 denote the horizontal and vertical truncation in the image border. The statistics are shown in Fig. 7. More than half of the objects are partially or heavily occluded while the occlusion percentage of KITTI is be-

Figure 6. Top: The number of total annotated objects per image. Bottom: The number of total annotated 3D objects per image.

Figure 7. The occlusion and truncation distribution of coarse categories. Over half of the objects are partially or heavily occluded between 5% to 30%, which reflects the difficulty of our 3D perception dataset and task.

The ambiguity analysis. The roadside dataset has inherent ambiguity due to the adopted various cameras with different settings of camera specifications, mounting heights, the
pitch angles of viewpoint, so on. Thus we analyze the diversity distribution of settings in Fig. 8.

4. Task and Metrics
4.1. The task overview and metrics.

Task Overview. The monocular 3D perception task is to localize the objects given ambiguous images captured under various settings and scenes, including different camera specifications, viewpoints, and mounting positions.

Evaluation Metrics. There are multiple evaluation metrics for AD datasets. In KITTI [12], the 11-point Interpolated Average Precision metric (AP$_{R_{11}}$) and the average orientation similarity (AOS) are proposed to assess the localization and orientation performance. [35] suggests 40 recall positions instead of 11 positions for a more fair comparison. The nuScenes [5] consolidates the mean AP and five True Positive (TP) error types into the nuScenes detection score (NDS), including translation, scale, orientation, velocity, and attribute error types. Inspired by the AP metric, we adopt AP$_{R_{40}}$ [35], i.e.,

$$AP_{R} = \frac{1}{|R|} \sum_{r \in R} \max_{\rho (\hat{r})} \rho (\hat{r})$$  (1)

where $\rho (\hat{r})$ is the precision at a certain recall threshold $r \in \{1/40, 2/40, ..., 1\}$. To facilitate a comprehensive evaluation w.r.t. a certain factor such as orientation, we decouple the consolidated metric into several sub-metrics.

Average Ground Center Similarity. The distance between objects on the ground plane indicates the risk of collision, we hence compute the ground Euclidean distance by projecting the object center points onto the ground plane:

$$ACS = \frac{1}{|D|} \sum_{s \in D} \left(1 - \min \left(1, \frac{\Delta_s^C}{C_s}\right)\right)$$  (2)

where $D$ is set of true positive samples, $C_s$ is the norm of the GT ground center, $\Delta_s^C$ is the Euclidean distance between predicted ground center and GT ground center of sample $s$. $|D|$ is the total number of true positive objects.

Average Orientation Similarity. The Average Orientation Similarity (AOS) is introduced to measure how well the orientation is estimated, which is defined similarly as [38] by,

$$AOS = \frac{1}{|D|} \sum_{s \in D} \frac{1 + \cos (2 * \Delta_s^\omega)}{2}$$  (3)

where $\Delta_s^\omega$ is the angle difference of sample $s$, and $\cos (2 * \Delta_s^\omega)$ means that during evaluation, we don’t distinguish whether the head or tail of the object is facing the camera.

Average Area Similarity. We measure the ground occupancy of the prediction w.r.t the GT in term of area, where $\Delta_s^A$ is the absolute area difference and $A_s$ is the ground truth area.

$$AAS = \frac{1}{|D|} \sum_{s \in D} \left(1 - \min \left(1, \frac{\Delta_s^A}{A_s}\right)\right)$$  (4)

Average Four Ground Points Distance and Similarity. We also compute the average distance of four ground vertices of the 3D bounding box (AGD), since it consolidates the location, orientation, and width/length together.

$$AGD = \frac{1}{|D|} \sum_{s \in D} \left(\frac{1}{K} \sum_{g=0}^{K-1} |s_g - \tilde{s}_g|\right)$$  (5)

where $\tilde{s}_g$ and $s_g$ are the $g$th predicted and GT ground points of smaple $s$, respectively. $K = 4$ is the total number of ground points. To be consistent with other similarity metrics, we define the AOS (ground points similarity) as:

$$AGS = \frac{1}{|D|} \sum_{s \in D} \left(1 - \min \left(1, \frac{1}{K} \sum_{g=0}^{K-1} |s_g - \tilde{s}_g| / |\hat{c}|\right)\right)$$  (6)

Assume $S = (ACS + AOS + AAS + AGS)/4$, we consolidate into Rope$\text{score}$ by reweighting the 3D AP and the proposed similarities metrics with $\omega_1 = 8$ and $\omega_2 = 2$.

$$\text{Rope} \text{score} = (\omega_1 * AP + \omega_2 * S) / (\omega_1 + \omega_2)$$  (7)

4.2. 3D Roadside Perception Task

As is illustrated in Sec. 1, due to the inherent ambiguity of roadside data caused by diverse camera specifications (various intrinsics and mounting positions, et.al.), the existing frontal-view monocular 3D object detection approaches can not be directly applied to the Rope3D Dataset. Hence
we make simple-and-effective attempts to alleviate the ambiguity problem by utilizing camera specifications and encoding the ground knowledge. Two modifications (early-fusion and deep-fusion) to incorporate the depth map of the ground plane with RGB image and two kinds (integrate and multi-gridded) of ground planes are made to alleviate the multi-focal ambiguity.

**Adaptations by leveraging ground planes.** We adopt the ground plane equation \(G(\alpha, \beta, \gamma, d)\) and camera intrinsic \(K^{3\times3}\) to generate the depth map \(D_G\) of the ground plane with the same size as the image.

\[
\begin{align*}
\{ Z \left[ x, y, 1 \right]^T &= K^{3\times3} \left[ X, Y, Z \right]^T \\
G^{4\times4} \left[ X, Y, Z, 1 \right]^T &= 0
\end{align*}
\]

where \([x, y]\) is the pixel in the image coordinates, \([X, Y, Z]\) is the corresponding 3d point in camera coordinate that lies on the ground plane. Thus the depth \(Z\) can be derived with the known 2d image points and the ground plane equation \(G\). We incorporate the ground depth map with the RGB appearance feature by early fusion and deep fusion. The first one is directly concatenating the depth map with the original RGB channels as input, and the second is adopting another siamese network for depth feature extraction and further weighted fusion of the two depth predictions. The performances of these two methods are similar and we hence only report the results by concatenation on the anchor-based M3D-RPN and the keypoint-based MonoDLE and MonoFlex approaches. We believe more sophisticated approaches might further improve the performance, which is out of the scope of this paper. In addition, two different formats of ground planes are attempted. One is to fit the entire ground within the visual field to a single plane, which is represented by the ground plane equation. Another is to divide the entire ground into multiple small grids, and each grid is represented by a ground equation.

5. Experiments

5.1. Experimental Setup

Our roadside 3D perception dataset contains 50k images, with the training and validation ratio set to 8:2. We offer two kinds of splitting the training and validation set. I: Homologous, for each scene we select 70% images and combine them for training, and leave all the rest images for validation. II: Heterologous, we select 80% of the cameras with the collected images for training and leave the remaining unseen 20% (different camera specifications) for validation, which can be used for validating the generalization ability of the monocular 3D object detection approaches.

**Implementation Detail.** (1) For M3D-RPN [3], we experiment on vanilla and improved approaches with ResNet34 [14] backbone, (2) Kinematic3D [4] is a monocular video-based 3D object detector with DenseNet121 [15] backbone, we only implement the first phase without video knowledge. (3) MonoDLE [25] is based on the anchor-free one stage detector CenterNet [51] with backbone DLA34 [48]. (4) MonoFlex [50] is a keypoint based method with modified DLA34 [48] backbone. The training image resolutions are adjusted to fit our dataset.

For objects that contain only 2D annotations without 3D annotations, we compute only losses on 2D attributes. For those objects having 3D labeling, the training loss weights of the 2D and 3D are both set to 1.

5.2. Main Results and Analysis

**Performance of vanilla and improved approaches.** The performances of monocular 3D detection approaches on the Rope3D Dataset are depicted in Table 2. Approaches with suffix \((G)\) denote that we customize the corresponding approach with ground plane function to reconnect the 3D locations and 2D projected points even when the optical axis of cameras is not parallel to the ground plane because of the pitch angle. The improved approach is noted with suffix \((D)\). We adopt the 3D \(AP_{1\text{m}}\) as well as the proposed metrics for evaluation. We find that most approaches have an obvious performance decline from homologous to the heterologous validation set. However, the performance drop is insignificant when applying a 3D detection model trained on the vehicle-view ONCE dataset to nuScenes dataset [27]. This phenomenon indicates the domain gap caused by various camera specifications and setting positions can not be ignored, which might be a distinguishable difference between vehicle-view and roadside view applications and should be taken carefully. By leveraging the depth map of ground planes, we observe an obvious improvement in most methods, even on the heterologous set where training and validating sets have different camera specifications.

**Performance of different ground plane formats.** We further analyze the performance by adopting two different formats of ground planes, i.e., by fitting the entire ground within the visual field to a single plane or by dividing the entire ground into multiple 5m×5m grids piecewisely. We carry out the experiment on KM3D [20], a method that predicts 2D keypoints and solves 3D position by minimizing the re-projection error. In other words, it relies on the differentiable geometric constraint to recover 3D location rather than direct prediction, which heavily depends on the accuracy of ground plane. As is shown in Table 3, KM3D-(GG), which takes advantage of gridded ground planes, shows better performance on most fine-grained categories. The main reason might be that the piecewisely-gridded planes better fit the actual ground plane.

**Performance of different ranges.** We further analyze the performance of the models within different ranges from 0 to 120m. As is shown in Table 4, with the depth range increases, most of the performances decrease, especially for
Table 2. Overall performance of the monocular 3D object detection approaches on the Rope3D Dataset with IoU = 0.5 and 0.7 under two train-val splitting settings: the homologous (I) and the heterologous (II). - (G) denotes adapting the ground plane, - (D) means using the depth map of ground. The abbr. in the branch column denotes: A: anchor-based, K: keypoint-based.

Table 3. Performance of KM3D with different ground plane formats on the Rope3D Dataset under two train-val splitting settings: the homologous (I) and the heterologous (II). - (G) denotes adapting the ground plane equation, - (GG) means using the gridded ground planes. IoU = 0.25 for non-motor vehicles and pedestrian, IoU = 0.5 for motor vehicles.

Table 4. The performance within different ranges on the heterologous (II) set. We set IoU = 0.5 for Car / Big Vehicle and IoU = 0.25 for Cyclist and Pedestrian, following KITTI’s hard-modality.

6. Conclusion

We propose the first high-diversity challenging roadside monocular 3D perception dataset - Rope3D. Rope3D is collected from the roadside view with joint 2D-3D annotations, making it unique from any previously released datasets and is particularly designed for the roadside 3D perception. Furthermore, we specially tailor the existing monocular 3D object detection approaches to the novel dataset, due to its unique viewpoint and inherent ambiguity lying in the various camera specifications and diverse road scenes. We hope to raise the attention to the special view - roadside view, so as to facilitate a safer and more intelligent autonomous driving system.

Ethical concerns and Limitation. To prevent from being utilized for illegal surveillance, all the images in the dataset are time-discrete and not allowed for tracking tasks. Note that all the sensitive information including license plates, human faces, names of bus stops, roads, and buildings are totally masked. The attempts for adapting the concurrent vehicle-view 3D detection approaches need further study.
References


[26] Yuxin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory


