AutoMine: An Unmanned Mine Dataset

Yuchen Li[1,2,3], Zixuan Li[1], Siyu Teng[2,3], Yu Zhang[4], Yuhang Zhou[4], Yuchang Zhu[4]
Dongpu Cao[1,5], Bin Tian[1,6], Yunfeng Ai[1,7], Zhe Xuanyuan[2], Long Chen[1,6*]

[1]Waytous [2]Beijing Normal University-HongKong Baptist University United International College
[3]HongKong Baptist University [4]Sun Yat-sen University [5]Tsinghua University
[6]Institute of Automation, Chinese Academy of Sciences [7]University of Chinese Academy of Sciences
liyuchen2016@hotmail.com zhexuanyuan@uic.edu.cn long.chen@ia.ac.cn

Abstract

Autonomous driving datasets have played an important role in validating the advancement of intelligent vehicle algorithms including localization, perception and prediction in academic areas. However, current existing datasets pay more attention to the structured urban road, which hampers the exploration on unstructured special scenarios. Moreover, the open-pit mine is one of the typical representatives for them. Therefore, we introduce the Autonomous driving dataset on the Mining scene (AutoMine) for positioning and perception tasks in this paper. The AutoMine is collected by multiple acquisition platforms including an SUV, a wide-body mining truck and an ordinary mining truck, depending on the actual mine operation scenarios. The dataset consists of 18*+ driving hours,* 18*K annotated lidar and image frames for 3D perception with various mines, time-of-the-day and weather conditions. The main contributions of the AutoMine dataset are as follows: 1.The first autonomous driving dataset for perception and localization in mine scenarios. 2.There are abundant dynamic obstacles of* 9 *degrees of freedom with large dimension difference (mining trucks and pedestrians) and extreme climatic conditions (the dust and snow) in the mining area. 3.Multi-platform acquisition strategies could capture mining data from multiple perspectives that fit the actual operation. More details can be found in our website(https://automine.cc).*

1. Introduction

Autonomous driving has received considerable attention in recent years and is becoming increasingly crucial in the field of artificial intelligence. It has been established that the public unmanned driving datasets play a role in validating algorithms. For example, some datasets , such as KITTI [9], Cityscapes [5], and A2D2 [10] have been stimulating researchers' interests. With the development of deep learning and other data-driven approaches, large volume datasets like ApolloScape [12] and Waymo Open [8] emerge to assist scholars to examine generalization ability of high complexity models. However, the pervasive autonomous driving datasets pursue massive redundant sensors and annotations which ignore the applicability on special scenes, including mines, ports, airports and so forth. Consequently, there is a lack of representative datasets to reveal extremely complicated challenges appearing on the above-ground mines. We publish our unmanned mine dataset (AutoMine) to fill this gap and attract attention to autonomous driving in mining scenarios.

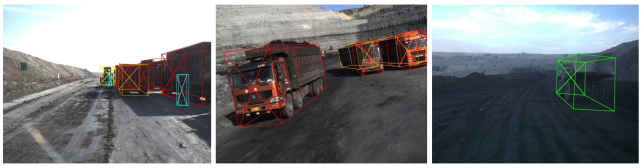

Figure 1. The environmental characteristics of the strip mine include rugged unstructured roads, the strong light exposure and dust. AutoMine annotates 3D objects with 9 degrees of freedom.

1.1. Characteristics

In comparison with the application of autonomous driving in urban roads, the high casualty rate and poor working conditions in mining environment make unmanned mines more urgently needed. Due to the difficulty of data acquisition, researchers can not investigate mining scenarios and assess model performances. AutoMine presents authentic mining data that are available for researchers to observe, measure the data, to locate and settle the problems, and to score and compare the approaches. The aim of this study is to facilitate the advancement and to contribute to the real-

*corresponding author

ization of unmanned mines.

To the author's knowledge, AutoMine is the first open-pit mines dataset for autonomous driving, embracing plenty of attributes. In particular, the unstructured roads are prevalent as shown in Fig. 1, which means the lack of the texture and fringe of the road surface, putting forward a great challenge to localization strategies based on pattern identification. Some frameworks associated with traffic lane, reflection intensity, and drivable areas detection covering structured avenues are likely to fail in mines. In addition, the number of dynamic targets and the richness of background are much less than those on urban datasets, imposing a great challenge on classification and object position. Therefore, how to detect potential loops and to enhance positioning and perception accuracy in scenarios lacking of abundant features are the unique values our dataset can provide to researchers.

We have discovered that the kinematic behavior of moving objects is affected considerably by the geological structure in open-pit mines. Hence, it is not appropriate to neglect the roll and pitch angle of each target on rugged and rough roads like most datasets. We precisely annotated 3D labels with 9 freedom including position, dimension and full-scale orientation as shown in Fig. 1. We also formulated a new 3D object detection metrics, aiming to accommodate the full degree of freedom outputs in mines. Through the investigation, we notice that intelligent vehicles also face tough weather and temperature challenges such as dust-storm, drizzle, heavy snow, as well as extremely cold and sunlight exposure. These types of adverse working conditions request higher demand to onboard sensors. Besides, the geometry discrepancies of dynamic targets (mining trucks and pedestrians) and apparent long-tail distribution of object types bring potential challenges to the perception missions.

Last but not least, it should be pointed out that trucks (including tractors and trailers), wide-body trucks and mining trucks are usual operating platforms in strip mines. However, there is a small proportion of civilian vehicles. In order to enhance the practical performance of the model, we apply a variety of platforms to collect mining data, including an SUV, a wide-body truck and a mining truck (Fig. 2). This adjustment might offer a more macroscopic perspective for scholars to understand the mining environment. Novel object detection metrics and repetition results of perception as well as localization are proposed and analyzed in our paper.

1.2. Related datasets

The publicity of various types of autonomous driving datasets have made a substantial contribution to the advancement in this area. KITTI [9] is a pioneering autonomous driving dataset, providing manifold computer vision tasks on urban roads in Karlsruhe. It is composed of

Figure 2. Three collection platforms are the Volkswagen Touareg SUV, the Tonly TLD65 mining wide-body transport truck and the Komatsu 930-4E mining truck from left to right.

22 driving scenes with more than 15K 3D annotations by a 64 lines lidar. Many researchers have been attracted to submit their testing outputs on KITTI website and got feedback with ranking. Cityscapes [5], BDD100K [32], Mapillary Vistas [20] have released plentiful data with segmentation masks. A*3D [22] enriches the sampling time and climate, specifically adds the dark night, rainy and snowy scenes into the visual set. NuScenes [2] is a large-scale autonomous driving dataset built by nuTonom with 40K calibrated frames and radar packets. Lyft Inc releases a Level 5 autonomous driving prediction dataset called Lyft L5 [11], containing more than 1K recorded driving hours with 55K 3D labeled boxes.

Some automobile manufacturers publish datasets collected by their vehicles, including H3D [21], A2D2 [10] and the Ford Dataset [1]. H3D is offered by Honda Inc including 1.1M labels with a complete 360-degree lidar in the San Francisco Bay. Audi's dataset A2D2 [10] involves 2D semantic segmentation, 3D point cloud classification, 3D border detection and bus control tasks. Ford discloses a luxuriant dataset (approximately 1.8T), Ford Dataset [1], covering 1K scenes, incorporating diverse seasons and constructed 3D maps.

Waymo [8] receives eminent popularity and outstanding reputation in the self-driving community, containing $1,150$ scenes with nearly 12 million 3D boxes on point cloud and 12 million 2D annotation boxes on images. Besides, ApolloScape [12] has been continuously updating, which now has released 147K+ labeled frames, including 100K high-resolution images with pixel-by-pixel semantically segmented information. Comma2k19 [24] is a highway dataset covering 33+ commuting hours on California's 280 expressways. The ONCE dataset [18] is made up of 144 driving hours with more than 1 million lidar scenes and 7 million corresponding camera pictures. Comprehensive comparisons between the AutoMine and other autonomous driving datasets are characterized in Tab. 1.

2. The AutoMine dataset

Here, we describe our dataset in detail by dividing it into several units, mining environment, collection platform, sensor configuration and synchronization, task partitioning, annotation statistics, localization and detection.

	Scenes	Time	Frame	Location	Road	3D-boxes	Night	Rain/Dust	Classes	9-freedom	Platform	Attribute
KITTI [9]	22	1.5	15K	Yes	Str	200K	Non	Non/Non	8	Non	Single	Urban
Cityscapes [5]	-	-	25K	Non	Str	-	Yes	Yes/Non	23	Yes	Single	Urban
nuScenes [2]	1K	5.5	40K	Yes	Str	330K	Yes	Yes/Non	23	Non	Single	Urban
A2D2 [10]	-	-	12K	Yes	Str	-	Non	Yes/Non	14	Non	Single	HW/Ur
Lyft L5 [11]	170K	1K	-	Yes	Str	-	Yes	Yes/Non	14	Non	Single	Urban
A*3D [22]	-	55	39K	Yes	Str	230K	Yes	Yes/Non	7	Non	Single	Urban
ApolloScape [12]	-	100	144K	Yes	Str	70K	Yes	Yes/Non	35	Non	Single	Urban
BDD100K [32]	100K	1K	100K	Non	Str	-	Yes	Yes/Non	10	Non	-	Urban
H3D [21]	160	0.77	27K	Non	Str	1.1M	Non	Non/Non	8	Non	Single	Urban
Argoverse [4]	113	0.6	22K	Yes	Str	993K	Yes	Yes/Non	15	Non	Single	Urban
Mapillary Vistas [20]	-	-	25K	Non	Str	-	Yes	Yes/Yes	66	Non	Multi	Urban
Waymo Open [8]	1K	10	200K	Yes	Str	12M	Yes	Yes/Yes	4	Non	Single	Urban
Comma2k19 [24]	1K	5.5	200K	Yes	Str	12M	Yes	Yes/Non	4	Yes	Multi	HW
Ford Dataset [1]	1K	5.5	200K	Yes	Str	12M	Yes	Yes/Non	4	Non	Single	Urban
PandaSet [30]	103	-	16K	Yes	Str	-	Yes	Non/Non	28	Non	Single	Urban
ONCE [18]	-	5.5	1M	Yes	Str	417K	Yes	Yes/Non	5	Non	Single	Urban
AutoMine	70	6.0	18K	Yes	**Unstr**	90K	Yes	Yes/Yes	9	Yes	Multi	**Mine**

Table 1. Comparisons with other public autonomous driving datasets. HW/Ur represents HighWay/Urban.

Environment We captured 18+ driving hours' data with 70 scenarios, utilizing three acquisition platforms (an SUV, a wide-body truck and a mining truck) at five strip mining sites in Inner Mongolia and Shaanxi province of China. Due to the scarcity of road features in mining environment and specific requirements, we require the route of the platforms to contain at least one global or partial loop of mining roads. In this way, we expect feature retrieval based localization algorithms with loop closure detection can be tested on our dataset to reduce the loss rate on unstructured roads. In practice, trucks commute between the excavating site and the dumping site alone the same route which further justifies the necessity of incorporating loops in the dataset, which elucidates our precondition is theoretically meaningful. Furthermore, unlike the climate conditions of urban, drivers sometimes experience extreme weather in mining areas, so we intentionally chose to collect data with distinctive weather conditions (snowstorms, dusty and sandstorms).

Collection platforms In order to satisfy the actual mining demand and the practicability of research, we choose three mobile acquisition platforms, which is one of the bright spots compared with other autonomous driving datasets with a single collection platform. We utilized a Touareg SUV, a Tonly TLD65 mining wide-body transport truck and a Komatsu 930-4E mining truck. The wide-body and mining truck are two of the practical operation vehicles in mines. These vehicles' appearance can be seen in Fig. 2 and their dimension values are in Tab. 3. Each collection vehicle contains at least one front lidar, an inertial navigation system, and two monocular cameras. There is growing evidence that the larger volume of trucks, the higher risk of self-driving operation has, so these types of acquisition platforms were equipped with multiple low-line lidars or light complement radars and millimeter wave radars. See Fig. 3

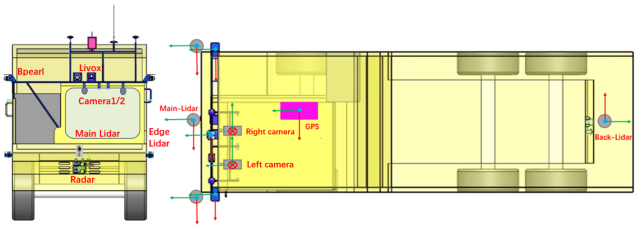

Figure 3. Sensor setup for the wide-body truck.

for sensor placement on the wide-body transport truck.

Sensor configuration Two FLIR industrial cameras were mounted on the top of three acquisition platforms with 55Hz capture frequency, $1/1.8$" anamorphic format, 2048×1536 resolution, 70 degrees Field of View (FOV). The Velodyne HDL-32E has 32 beams, 20Hz capture frequency, 360° horizontal FOV (only 180° valid), -30° $\sim$ +10° vertical scan scope, 70m range, ±2cm accuracy, up to 1.4M points per second and the Ouster-II-64 lidar has 64 beams with -7.9° $\sim$ +7.9° vertical angles, 20Hz, 150m range, up to 1.3M points per second. The inertial navigation system updates it's frequency in 10Hz. Tab. 2 shows models of other core sensors.

Sensor synchronization We adopted hard synchronization on the SUV to achieve sensor time calibration. To meet acceptable cross-modality data alignment between the lidar and the main camera sensor (left), the exposure of a camera was triggered when the main lidar swept across the center of the camera's FOV. The pose from left camera to front lidar (reference) was measured by mapping point cloud to corresponding pictures in the static situation. When the vehicle was moving, its speed was controlled below 20km/h and the fused data frequency was 5Hz. In addition, the main lidar is set as the reference of each platform, and the position as

	The SUV	The Wide-Body Truck	The Mining Truck
Main Lidar	Ouster OS-2-64 * 1	Velodyne HDL-32E * 1	Velodyne HDL-32E * 1
Camera	BFS-U3-31S4C-C *2	BFS-U3-31S4C-C *2	BFS-U3-31S4C-C *2
Inertial Navigation System	DAISCH * 1	DAISCH * 1	DAISCH * 1
Solid Laser	-	Livox MID-40	Livox MID-40
Edge Lidar	-	Velodyne VLP-16 * 3	Velodyne VLP-16 * 3
Blind Spot Lidar	-	RS-Bpearl * 3	RS-Bpearl * 3
Radar	-	ARS408-21 * 1	ARS408-21 * 3

Table 2. The sensors employed on three collection platforms.

Target Name	Dimension		
	Height	Width	Length
The SUV	1.65	1.74	4.69
The Wide-Body	3.95	3.47	9.10
The Mining Truck	7.40	8.72	15.03
Truck	4.87	3.80	11.9
Tractor	4.86	3.81	4.92
Trailer	4.86	3.81	9.85
Excavator	5.93	4.78	12.82
Pushdozer	3.37	4.45	8.31
Wide-Body Truck	4.06	3.55	9.53
Mining Truck	7.75	8.82	15.39
Civilian Vehicles	1.77	1.90	4.78
Pedestrian	1.81	0.75	0.89

Table 3. Our three different collection platforms' dimension and the average dimension value of dynamic objects.

well as pose of other sensors in this coordinate system are deduced by the calibration procedure to accomplish space synchronization.

Task partitioning At present, AutoMine supports two major autonomous driving tasks: localization and perception. Raw data is split into 70 independent streams, each of which lasts for $10 \sim 30$ minutes. Then by down sampling and semi-manually selection step, high quality data frames with more objects are served as the perception data. In the future, more enjoyable tasks such as segmentation and prediction will be taken into account in our dataset.

Data annotation For localization, the data includes position information (longitude, latitude and altitude) provided by the GPS and kinematic information (the speed and acceleration) supported by the inertial measure unit (IMU) of the collection vehicles. Perceptual annotation data consists of targets' 2D bounding boxes on image coordination, 3D properties with location (x,y,z), dimension (h,w,l), and rotation (roll,pitch,yaw), classification including trucks, tractors, trailers, wide-body trucks, mining trucks, excavators, pushdozers, civilian vehicles and pedestrians.

Annotation statistics We analyzed the relative elevation difference, data distribution and characteristics in the AutoMine. The duration of each piece of data is 15 minutes on average with an average of 50m in elevation difference, that reflects the prominent characteristics of roads in the mining area. More than 18K perception frames are annotated with 90K bounding boxes. The targets classification rotation distribution are shown in Fig. 4.

Localization Like most existing datasets, we offer the abundant vehicle location information based on the GPS and IMU. We recommend users to utilize the lidar-based, lidar-inertia SLAM(Simultaneous Localization and Mapping), pure vision, vision-inertia SLAM and fusion SLAM methods on our dataset. The purpose of whole process is to make full use of information to avoid localization loss and to boost the accurate rating on unstructured roads.

Detection There are 18K frames of data in the whole perception dataset (including multiple platforms). We divided it into training and testing set by 70% and 30%, and the annotated data provides roll and pitch angles, especially pitch angles, which often exists in the large uphill and downhill roads in mines. We hope that researchers can explore all the 9 DoF (degrees of freedom), with special attention to roll and pitch, which is essential for safe driving in strip mines rather than just the yaw angle. In addition, a truck with trailer is split in a head and a trailer to handle with the ambiguity when it turns a corner.

3. Task Metrics

3.1. Detection

The detection task for our AutoMine is divided into the 2D and 3D cases, both of which evaluate bounding boxes for 9 categories. To evaluate the inference results from the deep learning models, the outputs are sorted and categorized first by confidence and intersection values. Then the calculation strategy of 3D-AP is applied in accordance with based our newly proposed three-view perspectives, as opposing to the traditional Bird-Eye View only perspective in order to better evaluate the 3D detection models. Finally, we proposed a new metric called Detection Score (DS) with the intention of combining every individual score of 9 DoF.

Average Precision We employ the Average Precision (AP) metrics [9] [2] to assess 2D detection performances on the threshold by Intersection over Union (IoU) between predicting and ground truth bounding boxes on camera plane. To mitigate the impact of wide difference on dimension among different targets, We implement diverse IoU thresholds, $\{0.7, 0.6, 0.5\}$ for the trucks categories (including excavator, pushdozers and mining trucks), civilian vehicles and pedestrians. Predicted items with IoU under thresholds would be determined as false positive.

Moreover, We group object detection results in several sub-intervals by depth because we affirm remote objects in the mines are neither important nor easy to detect. The defined sub-intervals are $[0, 10]$, $[10, 20]$, $[20, 35]$, $[35, 60]$, $[60, inf]$. Consequently, separated by the recall rates r from 0 to 1 at the step 0.05, the AP could be obtained by summation of the weighted sub-interval area from drawing each precision-recall curve $p(r)$. α_{di} is the weight to indicate the importance of each depth intervals. Adding the weights is motivated by the analysis that numerous targets are concentrated at short range, so the evaluation favors the detection algorithm with accurate detecting capability to these targets within short ranges. Overall, the AP could be defined as:

$$AP = \sum_{d\in([0,10],[10,20]\cdots)} \alpha_{di} \int p(r)\mathrm{d}r \tag{1}$$

3D Average Precision There have been a number of autonomous driving studies involving Bird's Eye View (BEV) that have rendered effective on 3D perception. However, unlike other datasets only with single yaw orientation, our AutoMine's orientation information is composed of three classes, and these discrepancies exhibit complicated characteristics, requiring comprehensive measurement from multiple views besides BEV. For instance, Two completely overlapping boxes from the top view may have diverse pitch angles. Therefore, we use the mean IoU as the true positive metrics during 3D evaluation.

It must also be mentioned that when two boxes exist interaction on primary, top, and left views, it makes sense to calculate the 3D IoU score. Taking the pitch orientation observation into account, targets with large pitch angles on complex mining roads will generate upright rectangles which are much larger than the actual size, so contrary to straightway mapping 2D boxes from 3D, we firstly transfer the representation of the target's 3D attributes from the camera coordinate system to the local coordinate. Then the annotated pitch angle of each associated ground truth boxes would be eliminated that means various real boxes will become horizontal rectangles, and finally the coordinate system should be converted back after above elimination operation of the pitch angle. To meet the light computation requirement, rolling direction would not be converted, because the tiny rolling angle (less than 5 degrees generally) makes us conclude that the impact on the IoU is negligible. The mean IoU is defined as

$$mIoU = \frac{IoU_{ma} + IoU_{lef} + IoU_{top}}{3} \tag{2}$$

Eq. (2), IoU_{ma}, IoU_{lef} and IoU_{top} donate the IoU of the primary, left and top view respectively. According to Eq. (2), we utilize the same thresholds as 2D-AP mentioned above to distinguish positive samples from candidates. Then the AP_{3D} formula is shown as:

$$AP_{3D} = \sum_{d\in([0,10],[10,20]\cdots)} \alpha_{di} \int p(r)\mathrm{d}r \tag{3}$$

In contrast to the BEV score, the AP_{3D} in our metrics represents the combination result from multiple dimensions, especially the pitch angle from the left view. α_{di} is the weight as we mentioned in Eq. (1).

Besides the AP_{3D}, the independent metric of each degree of freedom in 3D space could assist researchers to explore partial advantages of the perception algorithm. Based on the matched positive samples, we define center distance, yaw, pitch, roll similarity and scale similarity. Considering center distance, since the dimension of the 3D bounding box from ground truth is basically similar to that from the associated corresponding prediction bounding box, and the distance between these boxes should be less than the summation by the range from the center point of each box to the boundary point. Take the ratio of the linear range between two centers to the summation as below:

$$Cd = 1 - \frac{1}{N\,(r_1 + r_2)} \sum_{d=1..N}^{N} d_i \tag{4}$$

d_i represents the rectilinear distance between two centers, and r_1, r_2 are the range between the center point of each 3D box to the vertex of it. Due to the reciprocal and subtraction operation, the center distance value is converted into $[0, 1]$. N donates the matched positive candidates number.

Yaw orientation is essential for 3D perception, Similar with yaw similarity mentioned in [9], it represents the normalized difference between the yaw angle of predicting and ground truth bounding boxes. Our yaw similarity only considers true positive samples.

$$YS = \frac{1}{N} \sum_{d=1..N}^{N} \frac{1 + cos(\triangle_{Yaw})}{2} \tag{5}$$

The formulations of pitch PS and roll RS indicators are similar to the yaw similarity, and we decouple them in order to verify the learning ability of different algorithms for various DoF in mines.

The size similarity evaluates the 3D dimensions including length, width, height of the true positive detection and g_x, p_x represent the ground truth as well as the prediction candidate.

$$SS = \frac{1}{3\mathrm{N}} \sum_{d=1..N}^{N} \sum_{x \in \{l,w,h\}} \min\left(\frac{p_x}{g_x}, \frac{g_x}{p_x}\right) \tag{6}$$

Detection Score Inspired by detection score in [5], we design a more comprehensive score through multiplying 3D-AP and the summation of these sub-items.

$$DS = AP_{3D} * \frac{3Cd + YS + PS + RS + 3SS}{9} \tag{7}$$

According to this formula, the DS receives the constraint by 3D-AP because the latter part of equation only takes true positive samples into account. Furthermore, on the basis of requirements of decoupled, we evenly distribute equivalent weights for the nine DoF and calculate the average.

3.2. Localization

In AutoMine localization task, we can divide it into visual and lidar localization, treating the GPS data as ground truth. In order to intuitively display the fault, we appraise the translational and rotational error.

ATE We use the Absolute Trajectory Error (ATE) to evaluate localization performance. ATE evaluates the absolute distance between the estimated and the ground truth trajectory, representing the global consistency of them. Because all tracks are in different coordinate frames, we need to map the estimated trajectory $P_{1:n}$ to the ground truth $Q_{1:n}$ through rigid transformation S. Therefore, the ATE at the time step i is defined as

$$F_i := {Q_i}^{-1} S P_i \tag{8}$$

We adopt the root mean squared error of the translation or rotation components at each moment.

$$ATE_{trans} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||trans\,(F_i)||_2^2} \tag{9}$$

RPE The Relative Pose Error (RPE) measures the discrepancy of the pose change within a fixed time interval Δ, which is suitable for estimating the drift of track. After aligning the timestamps, the relative pose error at time step i is the deviation between the estimated and the ground truth pose change within the time interval Δ. Therefore, the RPE at time step i is defined as

$$E_i := \left({Q_i}^{-1} Q_{i+\Delta}\right)^{-1} \left({P_i}^{-1} P_{i+\Delta}\right) \tag{10}$$

	2D-AP	YS	3D-AP	DS
PointPillar [14]	67.72	77.35	39.18	33.92
Second [31]	65.81	77.72	40.27	33.71
Second-IoU [31]	68.69	77.41	45.75	37.78
PointRCNN [27]	70.44	79.52	50.10	40.98
PointRCNN-IoU [27]	70.10	77.83	48.16	39.14
Part-Aˆ2-Free [28]	73.85	80.26	52.90	41.77
Part-Aˆ2-Anchor [28]	73.92	80.48	52.93	41.97
PV-RCNN [26]	77.39	76.02	54.59	46.87
Voxel R-CNN [6]	**81.14**	**88.93**	**55.37**	**47.11**
MonoGRNet [23]	55.81	64.02	4.66	4.09
SMOKE [17]	56.32	66.80	5.53	4.97
Stereo-RCNN [15]	60.49	70.57	8.92	7.47
YOLOStereo3D [16]	62.36	70.91	9.25	8.00

Table 4. The testing results of point cloud, monocular and binocular vision of trucks based perception algorithms on our dataset.

If the total number of pose sequences is n and the time interval is Δ, we can get $m = n - \triangle$ independent relative pose errors. We employ the root mean square error to compute the relative pose error as

$$RPE_{trans} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} ||trans\,(E_i)||_2^2} \tag{11}$$

4. Experiments

In this section we reproduce classical approaches for 3D object detection and localization tasks on the AutoMine. We also describe experimental results and attempt to analyze essential characteristics.

4.1. Baselines

First of all, we carried out a series of representative perception and localization algorithms for the baseline, categorized into lidar and visual based.

Lidar detection baseline For lidar based 3D detection tasks, we reimplemented two stages methods including PointRCNN [27], Part-Aˆ2 [28], PV-RCNN [26], Voxel R-CNN [6] and one stage such as PointPillar [14], Second [31] on OpenPCDet [29]. All of these strategies are prevalent in academia as well as industry. For evaluation on 2D-AP, predictions with confidence scores lower than 0.3 will be ignored, and we set the IoU threshold to 0.7 to identify positive/negative samples. We introduce yaw similarity (YS) which is similar with AOS indicators in KITTI, only considering matching boxes bias. In addition, the cosine similarity of the two additional angles is calculated, and the average angle cosine similarity consists of above direction elements by weighted combination. We define 3D-AP, generated by mIoU, the average Intersection over Union on the proposed

	Trans[m]		Rot[deg]	
	ATE	RPE	ATE	RPE
LOAM [33]	10.42	0.80	94.49	1.68
A-LOAM[1]	19.13	0.86	64.05	1.13
Lego-LOAM [25]	20.95	0.87	93.41	2.65
HDL [13]	11.86	0.94	76.28	1.42
ORB-SLAM2 [19]	23.06	3.01	115.09	5.05
ORB-SLAM3 [3]	24.75	2.14	117.64	5.11
DSO [7]	18.60	9.87	117.02	8.72

Table 5. The testing results of lidar-based and monocular vision localization algorithms in AutoMine.

three views (primary, top and left), to distinguish positive boxes from entire results. During the experiments, we refer the mean dimension in Tab. 3 to set anchor sizes.

The performance results for the lidar based algorithms are reported in Tab. 4, which only considered the truck category. It indicates that Voxel R-CNN outperforms other methods with 55.37%, 47.11% on 3D-AP and DS. Comparing with 95.11% in KITTI testing set for Voxel R-CNN, the 81.14% in 2D-AP suggests that the model's manifestation capability needs to be further improved on our dataset.

Monocular 3D detection baseline Recent years monocular 3D detection is favored by numerous scholars because of its low dependence on sensor. We employed MonoGRNet [23] and SMOKE [17] on AutoMine. According to Tab. 4, the unsatisfactory results (56.32% of 2D-AP and 5.53% of 3D-AP) of trucks from these methods illustrates a few challenges that cannot be tackled with monocular detection algorithms in mining areas.

Stereo 3D detection baseline Stereo 3D detection develops from binocular reconstruction, which extracts conjunct information adopting optical parallax volume. We used Stereo-RCNN [15] and YOLOStereo3D [16] as detection architectures. Experimental results in Tab. 4 demonstrate the baseline of Stereo-RCNN with 8.92% and 7.47% in 3D-AP and DS. Compared to monocular detection, the binocular detection improves up to 4% in 3D, however, it still requires further efforts and attention to achieve better performance.

Visual localization baselines In order to illustrate the properties of visual localization methods in AutoMine, we evaluated ORB-SLAM2 [19], ORB-SLAM [3] and DSO [7] in VO (visual odometry). We adjusted parameters to avoid loss of track in localization. Monocular SLAM has a considerable degree of scale drift, so we carried out scale alignment in the evaluation. As the experiments illustrates in Tab. 5, the baseline of ORB-SLAM2, ORB-SLAM3 and DSO are 23.06, 24.75 and 18.60 meters in ATE of translation. We conclude that the performance of them in AutoMine is much lower than that in KITTI, and the visual localization of AutoMine is more challenging.

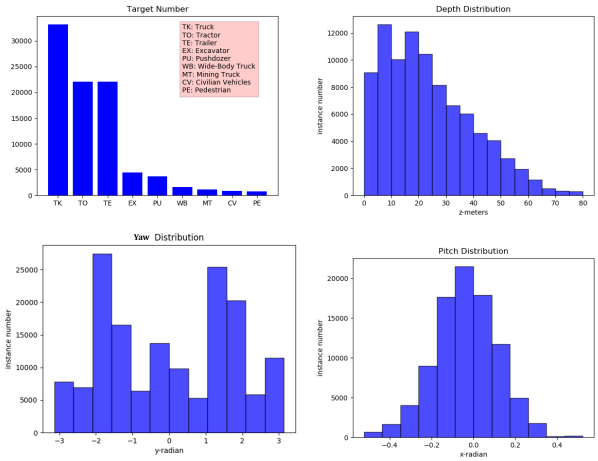

Figure 4. The distribution of categories, depth, yaw and pitch angle among whole targets in AutoMine.

Lidar localization baselines To demonstrate the properties of lidar localization algorithms on AutoMine, we reimplemented LOAM [33], A-LOAM[1], Lego-LOAM [25], HDL [13], and evaluated the performance in LO (lidar odometry). HDL took one lidar and GPS data as input, while other algorithms only used single lidar data. As shown in Tab. 5, four models perform poorly on AutoMine. Considering only the translational error, LOAM outperforms other strategies with 10.42, 0.8 meters on ATE and RPE. While, in the case of rotational error, A-LOAM has the better effect with values of 64.05 and 1.13 deg in ATE and RPE respectively, which is significantly ahead of other methods.

4.2. Analysis

Following the aforementioned benchmark, we analyze the crucial characteristics and difficulties that influence the indicators performance on our dataset.

Lidar VS. Camera The majority of the mining scenes in our dataset are monotonous on the point cloud from lidar and images from cameras. The feature extractors based on deep convolutional neural networks may produce confusion and generate unsatisfactory results. The aeolian landform and the dust raised by vehicles will create a large number of noise points. In addition, because of the particularity of mining roads, it is difficult to discern foreground objects on unstructured roads. It is clear that visual perception models are less effective than lidars, but under the mining environment, the gap has been widened. On the basis of whole dataset, we explore a number of images which are difficult to recognition like bright light exposure, giant halos, serious truncation especially the large size objects, and pixels

[1] https://github.com/HKUST-Aerial-Robotics/A-LOAM

	2D-AP	3D-AP	DS
VoxelRCNN	81.14	55.37	47.11
Replace Pit	82.75	57.74	47.29
Replace Yaw	81.26	55.50	47.13
Replace Dim	81.82	55.96	47.46
Excav only	75.41	52.10	46.22
Pedes only	50.17	37.58	32.54
Car only	62.31	41.73	37.99
Pedes/Truck	37.40/81.12	31.49/55.30	29.73/47.10

Table 6. We employed the ablation experiments on the angle and scale with replacing the truth to outputs, and different categories.

fuzzy from violent shaking. Except for cases mentioned above, two consecutive scenes from the dataset are captured at night and since the lack of illumination system in mining areas, it is extremely hard to identify objects in such black background images with only headlights. This is one of the most major reasons for weak visual detection accuracy.

Pitch & Roll Experimental algorithms only comprise 7 DoF for 3D detection, ignoring the pitch and roll, which seems to be reasonable in common rural roads, but these types of angle change dramatically when vehicles drive in uphill and downhill ways in mines. We replaced the annotated pitch orientation value to measured data for matched forecast objects and it is noteworthy that there are 2.37% and 0.18% increase of 3D-AP and DS respectively in Tab. 6. The benchmark model achieves a 3D-AP increase of 0.13% for substituting the rolling angle. This is consistent with the observation from the indistinctive distribution of rolling angles in Fig. 4.

Large size & Small size In order to estimate the effect of dimensions of targets, a series ablation experiments were conducted. Specific details are shown in Tab. 6. When the dimension was substituted with ground truth values, a small increase (0.59%, 0.35%) appears in the 3D-AP and DS. Moreover, when the performance of algorithms is assessed on the civilian vehicles and pedestrians, the accuracy decreases greatly, which is mainly related to the long-tail distribution of objects and sparse point cloud within targets in the dataset. In the multiple categories training, the detection accuracy of pedestrians further decreases because of inappropriate voxel resolution. The behavior of the dimension difference makes us conclude that the perception model or architecture applying in mines requires improvement to be compatible with various objects scales.

Unstructured scenes The topography of the open-pit mine is constantly changing as the excavation progresses, the trajectory of operating vehicles is not fixed, it is unnecessary and impossible to construct tarmac roads in mines. Therefore, the majority of the path is composed of soft sand and gritty soil, which lacks obvious curbs, lanes and features. These types of roads have an adverse effect on majority localization methods requiring more similar feature points between adjacent frames such as ORB-SLAM2 [19] and ORB-SLAM3 [3]. That is one of the dominant factors for low feature matching rates and high cumulative errors. It is worthwhile mentioning that few localization method is validated on unstructured roads, which impedes people to find out problems in these scenarios. We outline a possible solution, utilizing map matching at first and updating long-term map constantly in localization.

Sparse point clouds As shown in Tab. 5, the performances for lidar-based localization algorithms are not satisfactory. By analysing the data, we found that it is mainly due to the sparse point clouds in open-pit mines. The bumping and shaking exacerbate this undesirable situation when the main lidar is installed on the front of the trucks. Therefore, the combination of multiple lasers like edge lidars offers opportunities to mitigate the problem of data sparsity. In addition, we reckon that the ground modeling, normal vector analysis and other mathematical methods are supposed to play the key role in solving limited input data, and they would have more prospects in mining localization.

Multi-sensor fusion We believe that the multi-sensor fusion technology will boost the unmanned driving in mines. Our dataset currently involves vision and lidar data as input. However, the truck contains three extra low-line lidars, light complement and millimeter wave radars as shown in Tab. 2, and the data from these sensors can supplement the information of the blind area around the vehicle. Moreover, more ground points can be captured by these sensors, and as a result, they are likely to enrich input data of localization methods. Also, further features on targets' surface would be detected, thereby increasing the average precision. We aim to facilitate researchers by providing more data from various of sensors besides lidar and camera to tackle the driving safety issues in mines.

5. Conclusion

In order to solve the difficulties of autonomous driving vehicles on unmanned mines, we release the first autonomous driving dataset AutoMine for open-pit mines, which includes 3D object detection and localization tasks in mining areas with novel metrics, baselines and results. We expect our work to boost the research and development in unmanned mining.

Acknowledgements Our work are supported by the Key-Area Research and Development Program of Guangdong Province (2020B090921003), the National Key Research and Development Program of China under Grant (2018YFB1305002), and the research grant R201902 of UIC. We appreciate Waytous and Sun Yat-sen University for the hardware and software support at mines. Also, the annotated data are provided by the Institute of Automation, Chinese Academy of Sciences.

References

[1] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James R Mcbride. Ford multi-av seasonal dataset. *The International Journal of Robotics Research*, 2020. 2, 3

[2] H. Caesar, V Bankiti, A. H. Lang, S. Vora, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 5

[3] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 2021. 7, 8

[4] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 6

[6] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *arXiv preprint arXiv:2012.15712*, 2020. 6

[7] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 7

[8] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. Qi, and Y. Zhou. Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset. 2021. 1, 2, 3

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision Pattern Recognition*, 2012. 1, 2, 3, 5

[10] J. Geyer, Y. Kassahun, M. Mahmudi, X Ricou, and P. Schuberth. A2d2: Audi autonomous driving dataset. 2020. 1, 2, 3

[11] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, and P. Ondruska. One thousand and one hours: Self-driving motion prediction dataset. 2020. 2, 3

[12] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang. The apolloscape open dataset for autonomous driving and its application. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 1, 2, 3

[13] Kenji Koide, Jun Miura, and Emanuele Menegatti. A portable 3d lidar-based system for long-term and wide-area people behavior measurement. *IEEE Trans. Hum. Mach. Syst*, 2018. 7

[14] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[15] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo R-CNN based 3d object detection for autonomous driving. *CoRR*, abs/1902.09738, 2019. 6, 7

[16] Y. Liu, L. Wang, and M. Liu. Yolostereo3d: A step back to 2d for efficient stereo 3d detection. 2021. 6, 7

[17] Zechen Liu, Zizhang Wu, and Roland Tóth. SMOKE: single-stage monocular 3d object detection via keypoint estimation. *CoRR*, abs/2002.10111, 2020. 6, 7

[18] J. Mao, M. Niu, C. Jiang, H. Liang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, and J. Yu. One million scenes for autonomous driving: Once dataset. 2021. 2, 3

[19] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 7, 8

[20] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *IEEE International Conference on Computer Vision*, 2017. 2, 3

[21] 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. *Honda Research Institute 375 Ravendale Dr Suite B Mountain View CA 94043 USA*. 2, 3

[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. 2019. 2, 3

[23] Z. Qin, J. Wang, and Y. Lu. Monogrnet: A general framework for monocular 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2021. 6, 7

[24] H Schafer, E. Santana, A. Haden, and R. Biasini. A commute in data: The comma2k19 dataset. 2018. 2, 3

[25] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018. 7

[26] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. *CoRR*, abs/1912.13192, 2019. 6

[27] S. Shi, X. Wang, and H. Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[28] Shaoshuai Shi, Zhe Wang, Xiaogang Wang, and Hongsheng Li. Part-a^2 net: 3d part-aware and aggregation neural network for object detection from point cloud. *CoRR*, abs/1907.03670, 2019. 6

[29] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020. 6

[30] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, Yunlong Wang, and Diange Yang. Pandaset: Advanced sensor suite dataset for autonomous driving, 2021. 3

[31] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 6

[32] F Yu, Haofeng Chen, Xin Wang, W. Xian, Y. Chen, F Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. 2018. 2, 3

[33] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, volume 2, 2014. 7