Object Tracking Based on the Fusion of Roadside LiDAR and Camera Data

Shujian Wang, Rendong Pi, Jian Li, Xinming Guo, Youfu Lu, Tao Li, and Yuan Tian

***Abstract*— Tracking road users with high resolution is important for connected vehicles. Due to the complicated environments, tracking objects with a single sensor could not meet the requirements of high-resolution trajectories due to occlusions. How to acquire accurate and complete trajectories based on multisource data is a major challenge for researchers and engineers. This article developed a novel tracking method based on the fusion of roadside light detection and ranging (LiDAR) and camera. According to the relationship between the number of points and distance, the adaptive weight coefficient related to 3-D trajectory information was determined. The performance of the proposed method was evaluated at five selected sites. The proposed tracking method had high performance in terms of speed calculation, tracking range, the rate of object loss, and the repairing rate of disconnected trajectories. The proposed method can benefit many transportation areas, such as traffic volume counting, vehicle speed tracking, and traffic safety analysis.**

***Index Terms*— Adaptive weight, attention mechanism, object detection, object tracking, roadside light detection and ranging (LiDAR) and camera.**

I. Introduction

TRACKING road users is a critical technique for establishing an intelligent transportation system (ITS) [1]. Different from object detection, object tracking can be thought to be an extension of object detection. The trajectories of objects can benefit many applications, such as traffic congestion alleviation [2], road condition analysis [3], and traffic flow prediction [4]. To track objects accurately, many devices were employed, such as GPS [5], camera [6], radar [7], and light detection and ranging (LiDAR) [8]. However, most tracking methods were only developed based on a single sensor, such as a camera, resulting in low robustness in various environments due to occlusions, light, and so on [9]. Besides, the tracking range is limited due to sensor characteristics. With the traffic environment becoming more complex, detecting and tracking various road users are necessary by combining different sensors [10]. In general, these tracking-related devices can be deployed in drones [11], vehicles [10], and roadside infrastructure [8]. There are many studies aiming at object tracking based on onboard sensors [12], [13], that is, autonomous driving. However, only relying on onboard sensors cannot provide more accurate and sufficient information in complicated conditions. Therefore, how to track road users based on the fusion of different roadside sensors is a critical task in object tracking.

Generally, based on the number of tracked objects, the object tracking methods can be divided into single object tracking (SOT) [14] and multiobject tracking (MOT) [15]. The MOT plays an important role in object tracking with roadside sensors.

Frossard and Urtasun [16] proposed an end-to-end object tracking framework containing multiple independent networks. These networks were utilized to process some image and point cloud data tasks, such as object detection and information matching. In this framework, object detection and information matching modules were designed as deep network structures to speed up data processing. Zhang *et al.* [17] proposed a "sensor agnostic" framework that employed a "loss-coupled" approach to process point clouds and image data. Through the data fusion module and the adjacency matrix learning module, object detection and information matching processing of multisource data were realized, respectively, to achieve the purpose of real-time tracking of road targets. Luitesn *et al.* [18] solved the problem of target occlusion through 3-D reconstruction technology, improving the accuracy of target tracking. The proposed MOTS fusion framework consisted of two phases. However, these previous studies are mostly based on deep learning methods. With networks operating directly on point clouds and images, these methods are usually heavy and difficult to train, especially in the object tracking network. Furthermore, most of them are time-consuming and resource-intensive, which needs more computation resources.

To alleviate the computational burden, some studies employed statistical methods to achieve object tracking. As for object detection, the high-performance deep learning networks can be utilized to achieve it. Simon *et al.* [19] proposed the complexer-YOLO framework, which enables real-time detection and tracking road targets by decoupling images and

Manuscript received 11 May 2022; revised 3 August 2022; accepted 19 August 2022. Date of publication 26 August 2022; date of current version 12 September 2022. This work was supported in part by the National Nature Science Foundation of China under Grant 52002224, in part by the Major Scientific and Technological Innovation Project of Shandong Province under Grant 2020CXGC010118, in part by the National Nature Science Foundation of Jiangsu Province under Grant BK20200226, and in part by the Program of Science and Technology of Suzhou under Grant SYG202033. The Associate Editor coordinating the review process was Dr. Kok-Sing Lim. *(Corresponding author: Rendong Pi.)*

Shujian Wang is with the School of Civil Engineering, Shandong University, Jinan 250061, China, and also with Shandong Hi-Speed Construction Management Group Company Ltd., Jinan 250002, China (e-mail: shujian_wang@hotmail.com).

Rendong Pi was with the School of Qilu Transportation, Shandong University, Jinan 250061, China. He is now with the Autonomous Systems Laboratory, The Hong Kong Polytechnic University, Hong Kong (e-mail: rendong.pi@outlook.com).

Jian Li and Tao Li are with Shandong Hi-Speed Group Company Ltd., Jinan 250002, China (e-mail: jian_li1231@outlook.com; tao_li_123@hotmail.com).

Xinming Guo and Yuan Tian are with the School of Qilu Transportation, Shandong University, Jinan 250061, China (e-mail: 202115385@mail.sdu.edu.cn; yuantian@sdu.edu.cn).

Youfu Lu is with Shandong Hi-Speed Construction Management Group Company Ltd., Jinan 250002, China (e-mail: youfu.lu@outlook.com).

Digital Object Identifier 10.1109/TIM.2022.3201938

point cloud data. In the 3-D object detection stage, the 2-D semantic information extracted by the network was attached to the 3-D point cloud point by point, and then, the semantic point cloud was entered into the 3-D complex-YOLO to achieve 3-D object detection. To achieve real-time detection of targets, this framework proposed a Labeled Multi-Bernoulli Random Finite Sets Filter (LMB RFS) to decouple and separate target detection and tracking. In conclusion, these previous studies mostly employed onboard devices to achieve object detection and tracking. The developed methods were more suitable for autonomous driving, not the roadside system. Furthermore, as previous studies told, the object detection and tracking networks were usually united. In this way, the failure of one sensor can easily cause the failure of the whole system.

To solve the mentioned problems, we propose a tracking method based on the roadside LiDAR and camera data. Considering the role of the roadside tracking system in ITS, which means that the roadside system needs to exchange information between the road users and roadside infrastructures, the tracking method based on roadside sensors should be less computational resource than deep learning methods as mentioned above to ensure the instantaneity of the roadside system. In this article, the proposed tracking method is just based on mathematical derivation. This tracking method can present a good performance only by CPU reducing the need for computational resources greatly. Furthermore, to make the tracking system more robust and stable, we design the proposed tracking method to process camera and LiDAR data separately. This is a good strategy to guarantee the performance of the roadside system to avoid system failure.

Besides that, the fusion of different data can be divided into two types: early fusion [17], [19] and late fusion. The early fusion can provide more accuracy than late fusion. However, this method can increase the computational burden and make fault tolerance of the system poorer, which cannot meet the requirements of the roadside tracking system. Therefore, in this article, we employ the late fusion to combine the image and point cloud data.

The main contributions of this article are given as follows: 1) fusing the attention mechanism, the proposed Yolov5s-CoordAtt and PointRCNN-SENet perform better than the original models and 2) by studying the relationship between the trajectory information and the distance, we propose a novel tracking method by fusing 2-D and 3-D trajectory information. The proposed tracking method improved the tracking range and fixed the disconnected trajectories. The problem of object loss got ameliorated after fusing multisource trajectory information. The remainder of this article is organized as follows. Section II introduces the data collection, object detection, and tracking in detail. Section III describes the experimental results. Conclusions are summarized in Section IV.

II. Materials and Method

In this article, the process of acquiring the road users' trajectories can be divided into three parts, as shown in Fig. 1.

The first part is data collection describing the related devices and process of collecting data. The second part is object detection and tracking. An improved object detection method with attention mechanisms was proposed to detect and classify various road users. The Deepsort and AB3DMOT were employed for 2-D and 3-D object tracking based on the detection results. 2-D or 3-D object detection and tracking results were not interacted before fusing multisource trajectory information. Finally, a novel tracking method with adaptive weight was proposed to extract high-resolution and complete road users' trajectories by converging the 2-D and 3-D trajectory information.

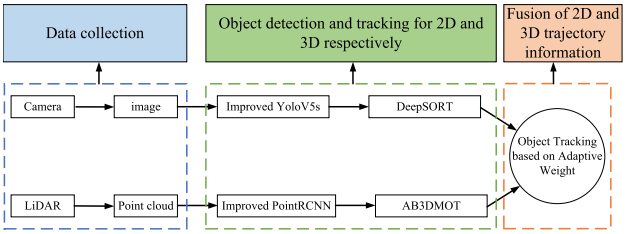

Fig. 1. Flowchart of acquiring object trajectories.

TABLE I
Camera Specifications

Indicator	Value
Product model number	HY1080
Scan FOV	90°
Image resolution	640×480 VGA MJPEG 800×600
Output image format	MJPEG YUY2
Working temperature	0~60°C
Working voltage	DC 5V

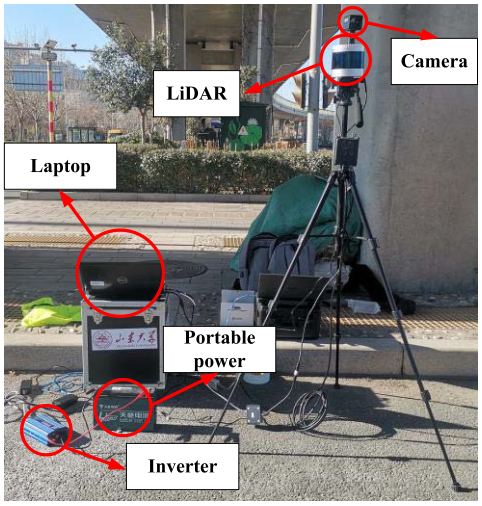

Fig. 2. Multisource data-collecting platform.

A. Data Collection

In this article, a LiDAR and camera were employed for multisource data collection. More detailed information about these devices is shown in Tables I and II.

The corresponding data-collecting platform has been developed with the LiDAR and camera, as mentioned above, as shown in Fig. 2. The laptop, portable power, and inverter are also the main components. 12-V dc was provided by the portable power (usually as a lead accumulator). The inverter

TABLE II
LiDAR SPECIFICATIONS

Indicator		Value
Product model number		LS-C32
Laser beams		32
Range		100~200m
Range resolution		±3cm
Laser wavelength		MJPEG YUY2
Scan FOV	Vertical direction	-16°~15°
	Horizontal direction	0°~360°
Angle resolution	Vertical direction	1°
	Horizontal direction	5Hz:0.09°/10Hz:0.18°/20Hz:0.36°
Rotating frequency		5Hz/10Hz/20Hz
Weight		1600g(standard)/1100g(light weight)

can increase the voltage from 12 to 220 V, meeting the requirements of LiDAR and laptop. As for the camera, it can be used when plugged into the laptop's USB port. The specific procedures of data collection are described as follows.

Step 1: Activate the driver of LiDAR and camera (a.launch file).

Step 2: Collect LiDAR and camera data by inputting the command "rosbag record."

Step 3: To avoid the data file becoming too large, stop recording data every 5 min.

B. Improved Object Detection Methods With Attention Mechanism

To extract road users' trajectories accurately and continuously, acquiring high-quality detection results is crucial. Attention mechanisms can make full use of data information under limited resources through various dimensions, such as space or channel. Generally, various attention mechanisms usually augment the important features and suppress nonimportant features. Due to their characteristics, there were many improved methods with attention mechanisms in computer vision [20] and natural language [21] in recent years. In this article, four attention mechanisms were employed to improve object detection methods' performance. These four attention mechanisms are SELayer [22], Ecalayer [23], CBAM [24], and CoordAtt [25]. The detailed elaboration for improved 2-D and 3-D object detection methods is described as follows.

As for 2-D object detection, YoloV5s were employed as the baseline in this article. To improve the network's performance in extracting features of the object, the attention mechanism was added to the backbone of the model. The architecture of the origin and improved YoloV5s backbone is shown in Fig. 3. The attention module was added to the four places in the backbone of YoloV5s, as shown in Fig. 3(b). In addition,

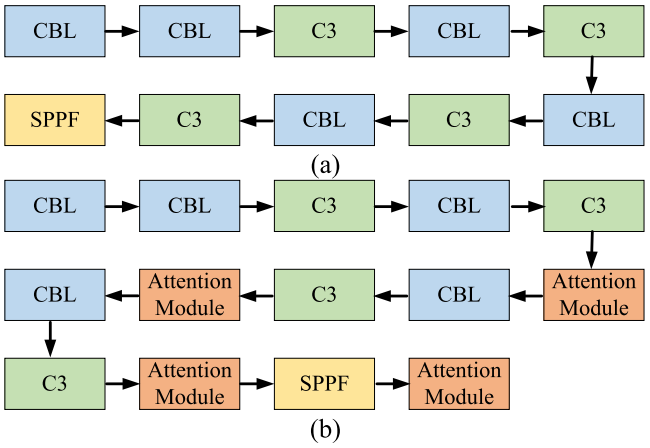

Fig. 3. Structure of (a) origin and (b) improved YoloV5s backbone.

the number of layers for these attention modules was the same as their upper module.

As shown in Fig. 3(b), four locations were selected to add the attention modules. The first two locations are the front of the P4 and P5 layers, respectively. The image features extracted by P4 and P5 were input into the detection head for feature fusion, essentially determining the prediction result of the object detection head. The third location is the front of SPPF, transforming arbitrary-size feature maps into fixed-size vectors. The key features of image feature vectors can be better retained by adding an attention module to the front of SPPF. The last location is at the end of Yolov5s backbone. The image features can be further captured by adding an attention module to the end of the models.

As for 3-D object detection, PointRCNN [26] was employed as the baseline in this article. The Openpcdet was employed to acquire the detection results of point clouds. The structure of PointRCNN in Openpcdet was divided into four parts: Backbone 3-D, Backbone 2D, Dense Head, and RoI Head. The attention module was added to the RoI Head to augment the network's ability to extract the global features of point clouds. The architectures of origin and improved PointRCNN are shown in Fig. 4.

According to the previous research, the local and global features were the critical factors for object detection with point cloud data. However, the characteristics of the point cloud easily vary with the distance, the shape of the object, and so on, and the local features of the point cloud were challenging to capture. The global features present more stability than the local features. Therefore, object detection performance can be enhanced by promoting the ability to extract the global features of point cloud data. As shown in Fig. 4(b), the attention module was added to the RoI Head to promote the performance of PointRCNN. The attention module can augment the ability to capture the high-dimension global features after point cloud region pooling.

It is noted that the attention module is shown in Fig. 3(b), and Fig. 4(b) meant adding SENet, ECANet, CBAM, and CoordAtt into the origin models.

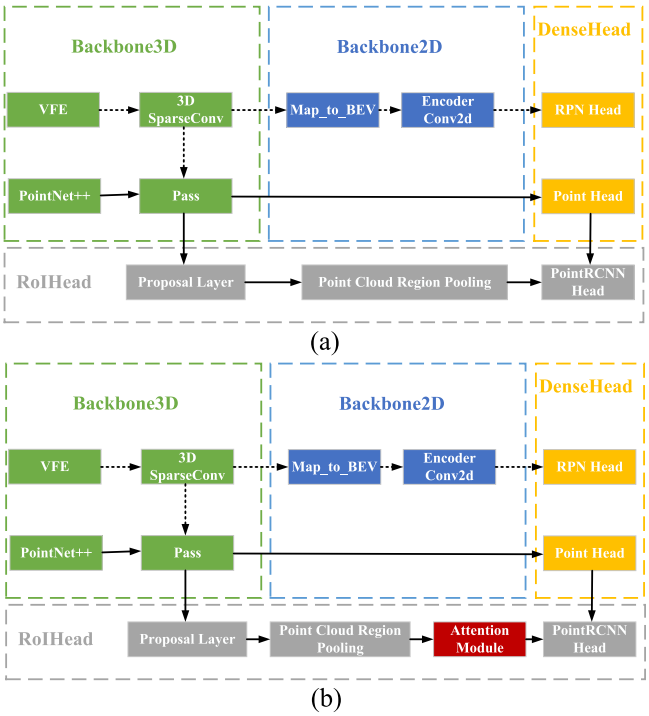

Fig. 4. Structure of (a) origin and (b) improved PointRCNN.

C. Object Tracking Method Based on Adaptive Weight

Based on the detection results, the object trajectories usually can be acquired by the Kalman filter [27] and the Hungarian algorithm [28], as demonstrated in many studies [29], [30]. A novel object tracking method was proposed in this article to fuse the 2-D and 3-D trajectories wholly and accurately. In fact, before employing the proposed object tracking method to fuse the multisource trajectory information, DeepSORT [30] and AB3DMOT [32] were utilized to extract the 2-D and 3-D trajectories, respectively.

Generally, the extracted trajectories can be divided into 2-D and 3-D trajectories based on the data source. In this article, track_{2D} and track_{3D} can be used to represent different trajectory information. The fusion of 2-D and 3-D trajectories can be denoted as

$$\text{track}_{\text{final}} = f(\text{track}_{2D}, \text{track}_{3D}, \alpha, \beta) \tag{1}$$

α and β, as the weight coefficients, are used to identify the proportion of 2-D and 3-D trajectory information in the final trajectory. According to the previous studies, the number of point clouds can decrease as the distance between objects and LiDAR becomes large. The density of point clouds showed the same pattern. The quality of point clouds reduces dramatically when the object is away from LiDAR [33]. However, the lower quality of point clouds can result in the failure of object detection and tracking. As for the detection and tracking methods related to images, they were less sensitive to distance. Even though the object was far from the camera, meaning that the object can only be identified among a few pixels in the image, the object can still be detected and tracked. In conclusion, compared to 2-D detection methods, the performance of 3-D

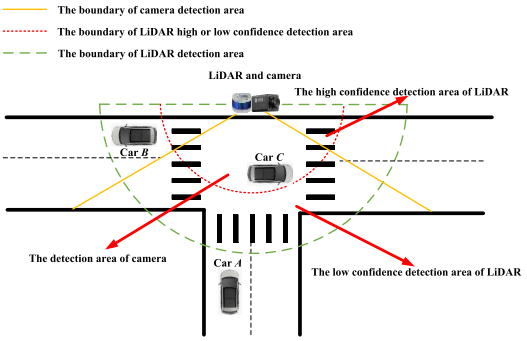

Fig. 5. Schematic of LiDAR and camera effective detection area.

can be easily affected by the distance between the object and data-collecting devices.

Taking the T-intersection as an example, the difference between the effective object detection and tracking range of LiDAR and the camera is shown in Fig. 6. The distance between the object and LiDAR plays an essential role in 3-D detection results. The confidence scores were always utilized to indicate the reliability of the object detection results. The effective detection area of LiDAR can be divided into two parts: the low- and high-confidence detection areas of LiDAR, as shown in Fig. 5.

There were three situations according to the distance between the object and data-collecting devices. The first situation is like Car A, which was not in the detection area of LiDAR but the detection of the camera. Under this condition, only the 2-D trajectory information can be acquired. Therefore, the final trajectory can be denoted as $\text{track}_{\text{final}} = \text{track}_{2D}$. The second situation is like Car B. The object was in the blind area of the camera and only can be detected by LiDAR. In this situation, the trajectory information of Car B can be represented by $\text{track}_{\text{final}} = \text{track}_{3D}$. If the object was Car C, it could be detected by LiDAR and the camera at the same time. The trajectory can be represented by the weighted average of 2-D and 3-D trajectory information, defined as follows:

$$\text{track}_{\text{final}} = (\alpha \cdot \text{track}_{2D} + \beta \cdot \text{track}_{3D})/(\alpha + \beta). \tag{2}$$

In conclusion, the true trajectory of the object based on the distance can be defined as follows:

$$\text{track}_{\text{final}} = \begin{cases} \text{track}_{2D}, & d > D_{\text{LiDAR}} \\ (\alpha \cdot \text{track}_{2D} + \beta \cdot \text{track}_{3D})/(\alpha + \beta), & d \leqslant D_{\text{LiDAR}} \end{cases} \tag{3}$$

where D_{LiDAR} is the maximum of LiDAR's effective tracking distance. track_{2D} denotes the 2-D tracking results by employing the DeepSORT, including the number of frames, the object category, the 2-D bounding boxes, and the tracking ID. track_{3D}, composed of the number of frames, the object category, the 3-D bounding boxes, the 2-D bounding boxes, and the tracking ID, can be acquired by AB3DMOT. In addition, the specific expression of the 2-D bounding box is (x_1, y_1, x_2, y_2), and the 3-D bounding box is $(h, w, l, x, y, z, \text{rot\_}y, \alpha)$, as shown in Fig. 6.

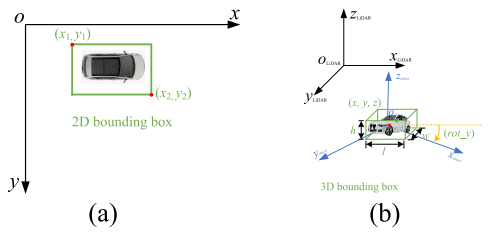

Fig. 6. Schematic of (a) 2-D and (b) 3-D bounding boxes.

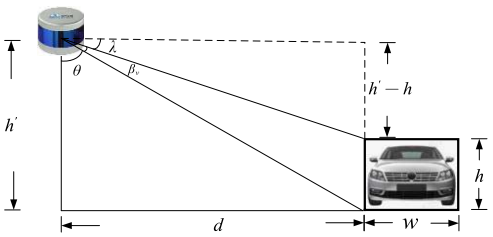

Fig. 7. Schematic of calculating the number of point clouds in the cross section of the object.

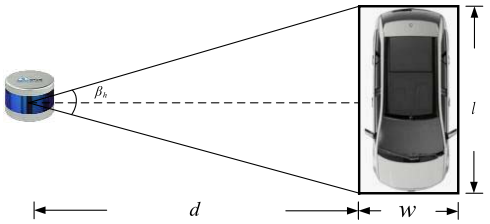

Fig. 8. Schematic of calculating the number of point clouds in the longitudinal section of the object.

As shown in (3), the critical task for fusing the 2-D and 3-D trajectory information is to determine the value of α and β. As described in Fig. 5, the 2-D and 3-D trajectory information has a significant difference from the perspective of distance. Specifically, the 2-D trajectory information is insensitive to the distance. Whether the object is at a far or close distance, it can both be detected and tracked accurately. Due to that the quality of the point cloud has dramatically decreased with the distance, the object represented by the point cloud can usually be detected and tracked at a closer distance than 2-D. The result of 3-D trajectory information is sensitive to distance. Therefore, we set the value of α and β based on the relationship between the trajectory information and distance in this article. Due to the difference in the distance sensitivity of 2-D and 3-D trajectory information, we suppose the value of α to be a fixed number 1, and the value of β was determined by the relationship between the number of point clouds and distance. Then, (3) can be transformed into the following equation:

$$\text{track}_{\text{final}} = \begin{cases} \text{track}_{2D}, & d > D_{\text{LiDAR}} \\ (1 \cdot \text{track}_{2D} + \beta \cdot \text{track}_{3D})/(1+\beta), & d \leqslant D_{\text{LiDAR}}. \end{cases} \tag{4}$$

According to the demonstrations mentioned above, the quality of 3-D detection and tracking results had a strong relationship with the distance. In addition, the distance had a significant impact on the number of point clouds. Therefore, the relationship between the weight coefficient of 3-D trajectory information, distance, and the number of point clouds can be denoted as follows:

$$\beta = f(d), \quad N_{\text{points}} = f(d) \tag{5}$$

where means the distance between the object and LiDAR, and is the number of point clouds. The connections between β and d can be established through the relationship between N_{points} and d.

The detected object can be supposed to be rectangular. Its length, width, and height were identified as l, w, and h. The installment height of LiDAR was h'. Due to the different angle resolutions of LiDAR in the horizontal and vertical directions, the number of point clouds should be considered from two directions, respectively.

As shown in Fig. 7, the angle can be calculated by (6) and (7) according to the geometrical relationship

$$\tan\lambda = \frac{h'-h}{d} \rightarrow \lambda = \arctan\left(\frac{h'-h}{d}\right) \tag{6}$$

$$\tan\theta = \frac{d}{h'} \rightarrow \theta = \arctan\left(\frac{d}{h'}\right). \tag{7}$$

The angle between the two laser beams can be calculated as follows:

$$\beta_v = \frac{\pi}{2} - \arctan\left(\frac{h'-h}{d}\right) - \arctan\left(\frac{d}{h'}\right). \tag{8}$$

Finally, the number of point clouds in the object's cross section can be represented by the following equation:

$$N_v = \beta_v / \mu_v \tag{9}$$

where μ_v is the angle resolution in the vertical direction.

As for the calculation of the number of point clouds in the horizontal direction, the relationship between the LiDAR and the object is shown in Fig. 8.

According to the geometrical relationship shown in Fig. 8, the calculation of β_h can be represented by the following equation:

$$\tan\frac{\beta_h}{2} = \frac{l}{2}\Big/ d \rightarrow \beta_h = 2\arctan\left(\frac{l}{2d}\right). \tag{10}$$

The calculation of the number of point clouds can be denoted as follows:

$$N_h = \beta_h / \mu_h. \tag{11}$$

In conclusion, the number of overall point clouds for an object was defined as follows:

$$N_{\text{points}} = N_v \cdot N_h = \frac{\frac{\pi}{2} - \arctan\left(\frac{h'-h}{d}\right) - \arctan\left(\frac{d}{h'}\right)}{\mu_v} \cdot \frac{2\arctan\left(\frac{l}{2d}\right)}{\mu_h}. \tag{12}$$

Furthermore, the number of point clouds for the rear of the vehicle can be calculated by the following equation:

$$N_{\text{points}} = N_v \cdot N_h = \frac{\frac{\pi}{2} - \arctan\left(\frac{h'-h}{d}\right) - \arctan\left(\frac{d}{h'}\right)}{\mu_v} \cdot \frac{2\arctan\left(\frac{w}{2d}\right)}{\mu_h}. \tag{13}$$

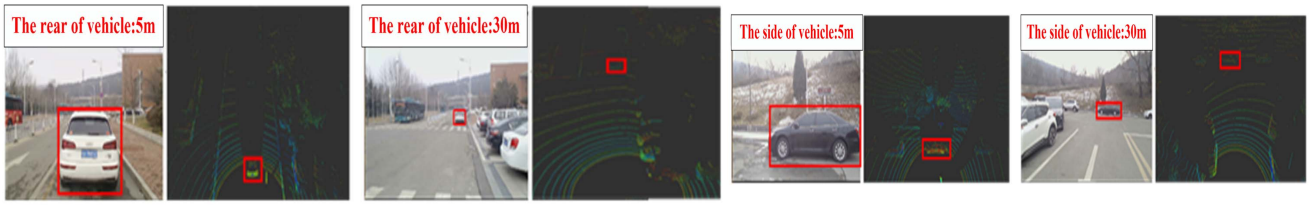

Fig. 9. Schematic of the validation experiment.

TABLE III
NUMBER OF POINT CLOUDS ACQUIRED BY THEORETICAL DERIVATION AND EXPERIMENT

Distance(m)		5	10	15	20	25	30	35	40
The side of the car	$N_{points\text{-}Experiment}$	1467	740	330	212	121	95	78	47
	$N_{points\text{-}Equ\ (11)}$	3405	901	404	228	146	101	74	57
The rear of the car	$N_{points\text{-}Experiment}$	740	380	174	91	56	42	28	23
	$N_{points\text{-}Equ\ (12)}$	1563	403	180	101	65	45	33	25

* $N_{points\text{-}Experiment}$ means N_{points} acquired by experiment, $N_{points\text{-}Equ\ (X)}$ means N_{points} obtained by Equation (X)

As shown in (12) and (13), it was easily found that the number of point clouds for an object was only determined by the installment height of LiDAR, the distance between the LiDAR and the object, the angle resolution of LiDAR, and so on. The relationship among them can be denoted as follows:

$$N_{\text{points}} = f\left(w, l, h, d, h', \mu_h, \mu_v\right). \tag{14}$$

However, the real shape of the car and other road users was not rectangular. The reality of (12) and (13) needs to be validated by the experiment. The specific procedure of the experiment is given as follows.

Step 1: Place LiDAR on a tripod.

Step 2: Collect the point cloud data of the side and rear of a vehicle at different distances.

Step 3: Count in the side and rear of the vehicle at different distances.

In addition, the distance was divided into five types: 5, 10, 15, 20, 25, 30, 35, and 40 m. Fig. 9 only shows some situations at different distances.

Based on (12) and (13) and the validation experiment, the results of under different distances are shown in Table III. In addition, the fitting curve of the number of point clouds calculated by (11) and (12) is shown in Fig. 10.

As shown in Fig. 10, there is a strong relationship between the theoretical value of N_{points} and the distance, denoted as (15) and (16). This finding was consistent with Song *et al.* [8]

$$f_{\text{side}}(d) = 1.2852 \times 10^4 \cdot e^{-2.7284 \times 10^{-1} \cdot d} + 1.1291 \times 10^2 \tag{15}$$

$$f_{\text{rear}}(d) = 6.0686 \times 10^3 \cdot e^{-2.7832 \times 10^{-1} \cdot d} + 5.0804 \times 10 \tag{16}$$

where d means the distance between the vehicle and LiDAR.

However, as shown in Table III, there is a significant difference between the experimental value and theoretical value of N_{points} when the object is close to the LiDAR. With the distance becoming large, the difference between the experimental and theoretical values in the number of point clouds gradually decreased.

The reason for this phenomenon can be summarized as follows.

1) The actual vehicle shape is not totally like rectangular. Therefore, the actual area scanned by the LiDAR is less than the area of a rectangular.
2) The laser can penetrate the windows. The point clouds belonging to this location are missing resulting in a further decreasing the number of the point cloud.

As mentioned above, the formula for calculating the number of point clouds needs to be revised based on the experimental result. In this article, the theoretical value was replaced by the experimental value at a distance of 5 and 10 m. Then, the formula for calculating the number of the point cloud can be transformed into (17) and (18). In addition, the fitting curves of these equations are shown in Fig. 11. R^2 told that there was a good fitness between the equation and data

$$f_{\text{side}}(d) = 2.6969 \times 10^3 \cdot e^{-1.2183 \times 10^{-1} \cdot d} + 2.3435 \times 10 \tag{17}$$

$$f_{\text{rear}}(d) = 1.4580 \times 10^3 \cdot e^{-1.3913 \times 10^{-1} \cdot d} + 1.8355 \times 10. \tag{18}$$

However, if we employ (17) and (18) to control the proportion of 3-D trajectory information in the final trajectory, this means that β was represented by the number of the point cloud. Due to the fact that the number of the point cloud is much larger than 1 (the value of α), the 3-D trajectory information will take a more significant account than 2-D. The 2-D trajectory information would get much less proportion in the final trajectory information. Therefore, the mix–max normalization method was employed to transform the range of the number of point clouds from 0 to 1. The corresponding normalization results

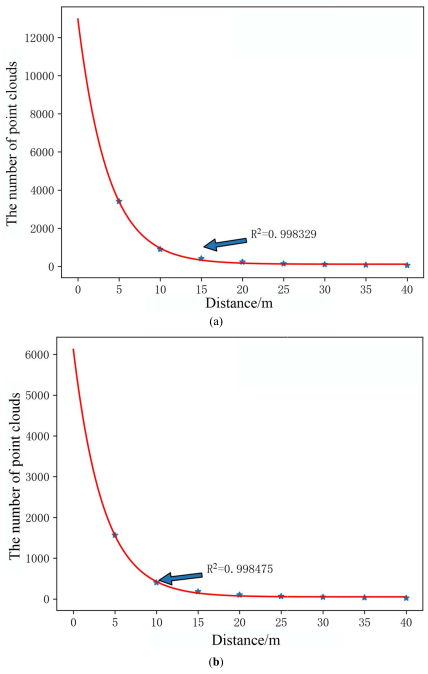

Fig. 10. Relationship between the distance and the number of point clouds. (a) Side of the car. (b) Rear of the car.

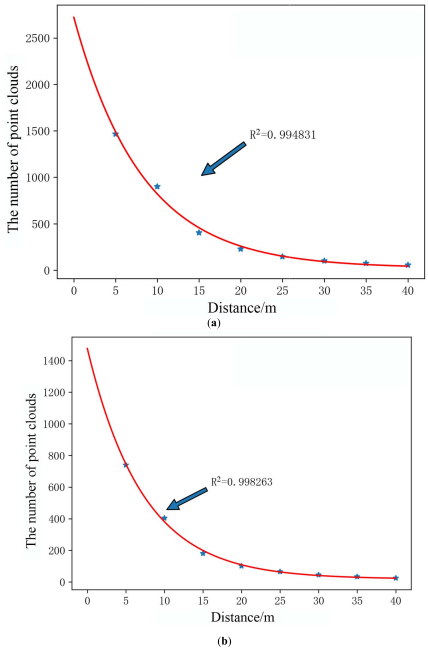

Fig. 11. Relationship between the distance and the number of point clouds. (a) Side of the car. (b) Rear of the car.

TABLE IV
VALUE OF β AT VARIOUS DISTANCES (AFTER REVISION)

Distance(m)	5	10	15	20	25	30	35	40
The side of the vehicle	1	0.5373	0.2857	0.1489	0.0744	0.0340	0.1197	0
The rear of the vehicle	1	0.4949	0.2429	0.1173	0.0546	0.0234	0.0078	0

are shown in Table IV and can be obtained by (19) and (20), as shown at the bottom of the page.

As shown in Fig. 12, there were three scenes in the view relationship between the vehicle and LiDAR. In the first scene, the number of point clouds can get the maximum. However, the number of point clouds can get the minimum in the third scene. To get a balance between the maximum and minimum, the weight coefficient can be calculated by the average value between the side and rear, as represented by the following equation:

$$\beta = \frac{f_{\text{sidefinal}}(d) + f_{\text{rearfinal}}(d)}{2}. \tag{21}$$

Finally, the value of β varying with the distance can be obtained, as shown in Table V. In addition, the precondition

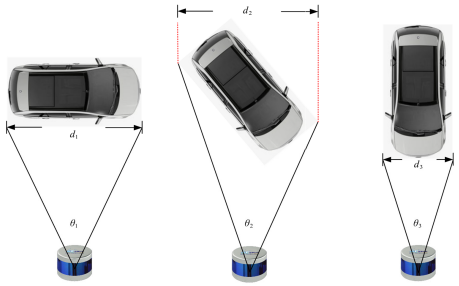

Fig. 12. View relationship between the vehicle and LiDAR.

TABLE V
FINAL VALUE OF β AT VARIOUS DISTANCES

Distance(m)	5	10	15	20	25	30	35	40
β	1	0.52	0.26	0.13	0.06	0.03	0.01	0

of fusing the 2-D and 3-D trajectory information was to link the same object's 2-D and 3-D information. In this article, the Hungarian algorithm was employed with the 2-D IoU as the loss function to converge the 2-D and 3-D information belonging to the same object. It is noted that calibration between the LiDAR and camera was based on Autoware.

$$f_{\text{sidefinal}}(d) = \frac{\left(2.6969 \times 10^3 \cdot e^{-1.2183 \times 10^{-1} \cdot d} + 2.3435 \times 10\right) - \min(f_{\text{side}}(d))}{\max(f_{\text{side}}(d)) - \min(f_{\text{side}}(d))} \tag{19}$$

$$f_{\text{rearfinal}}(d) = \frac{\left(1.4580 \times 10^3 \cdot e^{-1.3913 \times 10^{-1} \cdot d} + 1.8355 \times 10\right) - \min(f_{\text{rear}}(d))}{\max(f_{\text{rear}}(d)) - \min(f_{\text{rear}}(d))} \tag{20}$$

TABLE VI
VALUE OF β AT VARIOUS INSTALLMENT HEIGHTS OF LiDAR

Distance(m)		5	10	15	20	25	30	35	40
The height of Lidar was 1m	β	1	0.528	0.248	0.122	0.063	0.031	0.012	0
The height of Lidar was 1.5m	β	1	0.530	0.246	0.121	0.063	0.031	0.012	0
The height of Lidar was 2m	β	1	0.553	0.261	0.129	0.067	0.033	0.012	0

TABLE VII
OBJECT DETECTION PERFORMANCE IN TERMS OF THE INDICATOR BBOX

Modesl	Car			Cyclist			Pedestrian		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN	90.5758	89.4353	88.74	89.1802	76.7867	74.9714	73.5523	66.1911	62.1433
PointRCNN-CoordAtt	90.6615	89.4915	88.9964	89.3567	76.8559	75.0442	74.4089	66.9355	63.0299
PointRCNN-SE	90.027	87.3777	87.5843	97.2098	77.9501	75.0862	74.144	66.5688	62.705

TABLE VIII
OBJECT DETECTION PERFORMANCE IN TERMS OF THE INDICATOR BIRDS EYE VIEW (BEV)

Modesl	Car			Cyclist			Pedestrian		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN	89.8595	86.2811	80.0374	87.7141	73.9437	68.1489	67.4991	60.1117	53.1095
PointRCNN-CoordAtt	90.0633	87.6602	88.651	86.2377	71.9809	66.4139	68.9383	60.9939	54.7103
PointRCNN-SE	90.6031	89.3693	80.0353	88.2108	74.6208	72.5758	68.782	61.1962	54.6442

However, as shown in (14), the proposed method can be influenced by the installment height of LiDAR and the size of the object. Therefore, the impact of h' and the size of the object (ω, l, and h) were also studied.

Another car with a different size was selected, and the installment height of LiDAR was divided into three types: 1, 1.5, and 2 m. After employing the data processing methods mentioned above, we can get the value of β at various installment heights of LiDAR, as shown in Table VI.

From Table VI, it was easily found that, although h' may result in the increase or decrease in the number of the point cloud, the relationship between the number of point clouds and the distance obeys the same rule. From this respect, the proposed method is suitable for various installment heights of LiDAR. What is more, comparing the result shown in Table VI with the installment height of LiDAR being 1.5 m with Table IV, the size of the car would not have a huge impact on the value of β.

III. RESULTS AND DISCUSSION

A. Results of Various Object Detection Models

According to the results shown in Tables VII–X, these five models got the same performance when detecting the Car. However, as for the pedestrian and cyclist, the PointRCNN with adding SENet got better than other models. Specifically, the precision of Yolov5s-CoordAtt was higher than Yolov5s by 4.48%. From the perspective of point cloud object detection, PointRCNN-SENet almost had the same performance as PointRCNN in terms of Car. As for cyclists, PointRCNN-SENet was higher than the original model by 8% in terms of indicators average orientation similarity (AOS) and bbox. When detecting pedestrians, PointRCNN-SENte can also get better performance than PointRCNN. As shown in Table XI, Yolov5s with adding CoordAtt got the best performance among other models. Therefore, the CoordAtt and SENet were selected to add to Yolov5s and PointRCNN to obtain better object detection performance than origin models.

B. Design of Validation Experiment for Algorithm

In this article, the performance of the proposed method was evaluated from two aspects: accuracy and robustness. The accuracy of the proposed method can be evaluated by calculating the speed error between the baseline speed and extracted speed from the proposed method. As shown in Fig. 13, let a car drive in the straight two-way two-lane under different speeds. The multisource date-collecting platform was installed in the red star to collect the data as the car drives in lane A and lane B, respectively. The trajectory information can be extracted when the raw data were processed by the proposed method. In addition, the speed was divided into five types: 10, 20, 30, 40, and 50 km/h.

TABLE IX
OBJECT DETECTION PERFORMANCE IN TERMS OF THE INDICATOR3-D

Modesl	Car			Cyclist			Pedestrian		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN	87.9845	78.2125	76.8778	86.2524	71.2029	65.7126	61.5402	54.2067	49.7922
PointRCNN-CoordAtt	88.2703	78.7492	77.6747	84.9637	70.0036	65.1784	64.4323	56.238	51.6874
PointRCNN-SE	88.9696	78.1559	77.0894	86.9895	71.9057	67.0657	63.6939	55.9035	51.5747

TABLE X
OBJECT DETECTION PERFORMANCE IN TERMS OF THE INDICATOR AOS

Modesl	Car			Cyclist			Pedestrian		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN	90.56	89.34	88.54	89.09	76.33	74.5	70.84	63.32	59.1
PointRCNN-CoordAtt	90.59	89.28	88.47	89.27	76.28	74.38	71.38	63.79	59.76
PointRCNN-SE	90.65	89.37	88.79	97.11	77.53	75.05	71.58	63.73	59.64

TABLE XI
OBJECT DETECTION PERFORMANCE AMONG VARIOUS YOLOV5 MODELS

Models	Precision	Recall	mAP_0.5	mAP_0.5:0.95
YoloV5s	91.91%	86.59%	91.71%	0.6272
YoloV5s-SE	94.21%	83.41%	91.57%	0.6288
YoloV5s-ECA	95.76%	84.21%	92.07%	0.6388
YoloV5s-CBAM	95.58%	81.60%	90.40%	0.6014
YoloV5s-CoordAtt	96.39%	84.45%	92.81%	0.6441

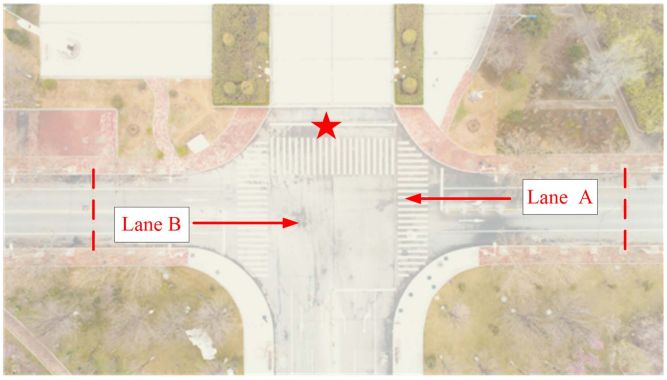

Fig. 13. Schematic of accuracy experiment for the algorithm.

According to, the calculation of 3-D speed was denoted as follows:

$$v_{3d} = \sqrt{(x_{a1} - x_{a2})^2 + (y_{a1} - y_{a2})^2 + (z_{a1} - z_{a2})^2}/T \quad (22)$$

where x_{a1}, y_{a1}, and z_{a1} mean the coordinate of the object in the frame T. In the frame $T + 1$, the coordinate of the object is x_{a2}, y_{a2}, and z_{a2}. In addition, T indicates the time interval between the adjacent frames. It is equal to the data-collecting frequency of LiDAR. In this article, T is 0.1 s.

The 2-D speed can be calculated through the method of image proportion coefficient. Given the focal length of the camera was easily acquired, the distance between the camera and the object can be calculated by the following equation:

$$\frac{D}{R} = \frac{f}{r} \quad (23)$$

where D is the distance between the object and camera, R means the real length of the object in the world, f is the focal length of the camera, and r indicates the size of the object in the image. The relationship between the unit pixel and unit length can be acquired based on the size of the image and camera, represented by the following equations:

$$D = \frac{R}{r} \times \frac{f \times W}{w} \quad (24)$$

$$D = \frac{R}{r} \times \frac{f \times H}{h} \quad (25)$$

where W and H are the width and the length of the image. The width and the length of the camera are w and h. Then, the image proportion coefficient can be defined as follows:

$$F_x = \frac{f \times W}{w} \quad (26)$$

$$F_y = \frac{f \times H}{h}. \quad (27)$$

Based on (26) and (27), (24) and (25) can be transformed as (28) and (29)

$$R = \frac{Dr}{F_x} \quad (28)$$

$$R = \frac{Dr}{F_y}. \quad (29)$$

Finally, the calculation of 2-D speed is defined as follows:

$$v_{2d} = R_{2d}/T. \quad (30)$$

To sufficiently evaluate the performance of the proposed object tracking method, five different sites were selected. Fig. 14 demonstrates the locations of selected sites. These locations are in the nonsignal intersection, intersection, roundabout,

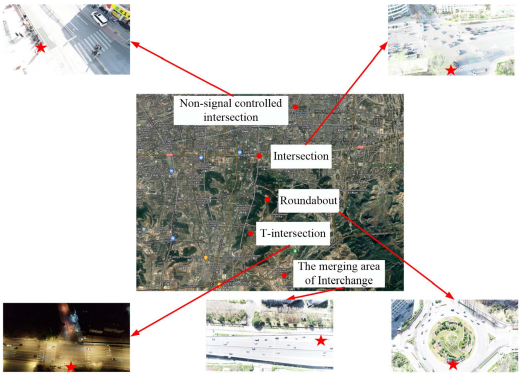

Fig. 14. Five different scenes employed for the performance evaluation of the proposed object tracking method.

T-intersection, and the merging area of interchange, respectively. The installation of the data-collecting platform was in the red star shown in Fig. 14. Generally, road users in these locations are prone to conflict. Acquiring high-resolution trajectories plays an important role to reduce the risk of conflicts among road users.

C. Performance Analysis of the Object Tracking Method Based on Adaptive Weigh

1) Accuracy Analysis: According to (22) and (30), we can get the 2-D and 3-D speeds, respectively. Then, the 2-D and 3-D trajectory information can be fused by the proposed tracking method. The speed error was employed as an indicator to evaluate the accuracy of the proposed tracking method, denoted by

$$S = \frac{|v_{\text{fusion}} - v_{\text{baseline}}|}{v_{\text{baseline}}}. \tag{31}$$

The results of speed under different situations are shown in Table XII. The speed error in the different situations was not more than 10%. The accuracy of the proposed method can get high performance at different speeds.

2) Robustness Analysis: The robustness of the proposed method was evaluated in three parts: the object tracking range, the object loss rate, and the repairing rate of disconnected trajectories.

As for the object tracking range, take the tracking results of the T-intersection as an example, as shown in Fig. 15. The 2-D and 3-D tracking results were visualized in Fig. 15(a) and (b), respectively. The object with a red rectangular box at the end of the branch road of the T-intersection was detected by the 2-D tracking method. However, the 3-D tracking method cannot track this object at the same distance. According to Section II-C, the number of points decreased with the distance. At such a distance, the number of point clouds was not enough to meet the requirements of the 3-D detection method to classify the type of object. Furthermore, the global and local features of the point clouds cannot be extracted sufficiently by the 3-D detection method. However, by employing the proposed method, the 2-D and 3-D trajectory information can be acquired in one frame at the same time, as shown in Fig. 15(c).

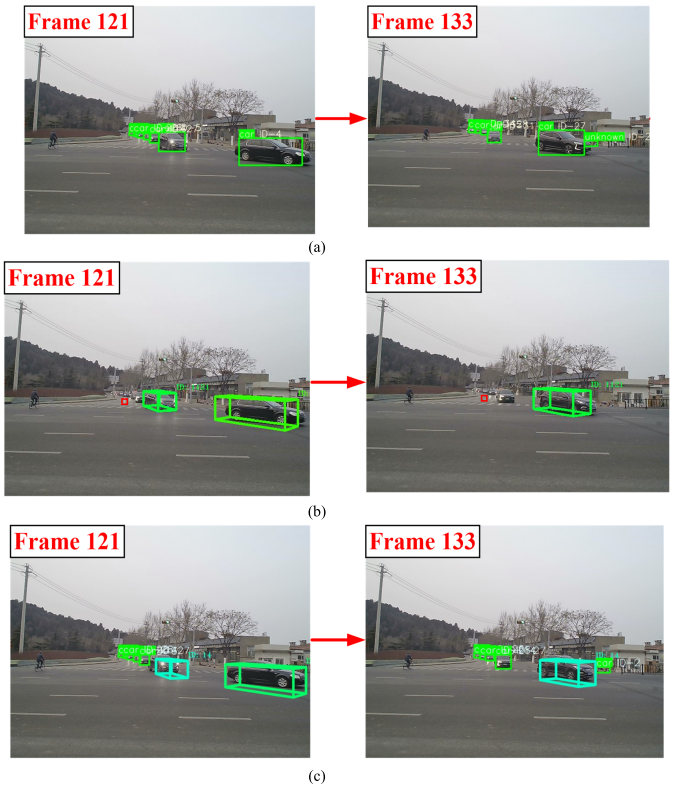

Fig. 15. Tracking results at the T-intersection. (a) 2-D tracking result. (b) 3-D tracking result. (c) Fusion tracking result.

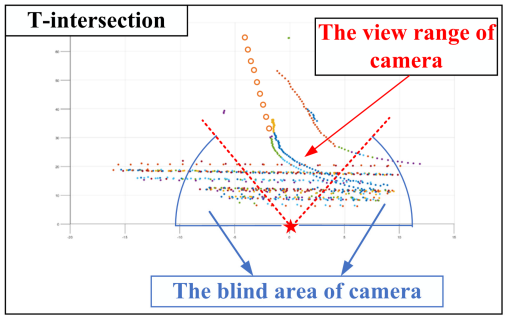

Fig. 16. Comparison of different tracking methods in the view range.

Compared to 2-D and 3-D tracking methods, the proposed method significantly improved the tracking range mainly through two aspects: the distance and view range. According to the tracking results shown in Fig. 15, the 2-D trajectory information can be acquired about 100 m away from the data-collecting platform. However, when the object was about 30 m away from the data-collecting devices, it can be detected and tracked. In conclusion, compared to the 3-D tracking method, the proposed method can extract the trajectory information at a far distance.

As for the view range, LiDAR can detect the object at the horizontal of 360°. However, the view range of the camera is only 90° according to Table I. As shown in Fig. 16, the area between two red dashed lines was the effective view range of the camera. These two blue fans were the blind areas of the camera. Due to the characteristics of LiDAR, the object in the blind area can be detected by LiDAR. The proposed method can extract the 2-D and 3-D information at the same time,

TABLE XII
RESULTS OF VEHICLE SPEED EXPERIMENT

Speed category	Frames										Average speed (km/h)	Baseline Speed (km/h)	Speed error
	1	2	3	4	5	6	7	8	9	10			
10A	10.2	10.6	9.7	11.9	10.8	11.6	10.2	8.9	9.2	9.8	10.2	10	2.9%
10B	11.7	12.4	11.8	11.8	12.4	9.8	10.6	9.8	8.9	10.3	10.9	10	9.5%
20A	21.9	22.1	20.3	21.5	20.8	21.4	19.8	19.7	20.8	22.8	21.1	20	5.6%
20B	22.8	21.7	20.4	20.6	19.8	19.7	22.3	22.4	19.8	20.1	20.9	20	4.8%
30A	37.9	35.2	33.2	28.9	29.1	32.4	33.7	30.6	29.4	30.8	32.1	30	7.07%
30B	29.6	30.2	31.5	32.8	35.4	34.6	31.2	29.7	30.8	30.9	31.6	30	5.57%
40A	39.7	37.1	38.6	35.4	34.5	40.1	39.7	41.2	39	39.7	38.5	40	3.75%
40B	37.8	39.4	38.4	48.6	44.9	48.7	42.3	41.9	40.6	35.6	41.8	40	4.55%
50A	47.6	48.3	45.9	46.8	47.6	46.8	48.3	49.7	49.2	48.6	47.8	50	4.24%
50B	52.9	51.6	54.3	56.1	50.2	49.8	49.1	51.1	51.3	50.6	51.7	50	3.4%

enhancing the view range greatly compared to the individual 2-D tracking methods.

Approximately 1200 frames of data collected from the selected five sites were used to summarize the trajectory information, respectively. The maximum tracking distance and the view range of object tracking can be obtained by the object coordinate in the trajectory information. The results are described in Table XIII.

As shown in Table XIII, the proposed method increased the tracking distance by 1.96, 1.45, 2.09, 1.99, and 1.16 times, respectively. Compared to the individual 3-D tracking method, the tracking distance improved by an average of 1.73 times. In the view range of tracking, compared to the individual 2-D tracking method, the proposed method can acquire the object in the blind area of the camera containing 3-D trajectory information.

The second part of the robustness was the object loss rate. Although improved 2-D and 3-D object detection methods with attention mechanisms were proposed in this article, the higher precision cannot indicate a good performance in capturing the object. Capturing the object as much as possible is the precondition for tracking multiobjects accurately. However, various environments can put different effects on the tracking results. It was necessary to evaluate the performance of object capturing at different sites by employing the object loss rate. In this article, we define the rate of object loss as the ratio of the number of object losses to the number of objects over a period of time. The rate of object loss can be denoted by

$$R_{\text{loss}} = \frac{N_{\text{object}}}{N_{\text{object loss}}}. \tag{32}$$

Similarly, a total of 1200 frames of data collected from the selected sites were used to summarize the trajectory information, respectively. Due to the significant difference in the detection results among various road users, vehicles, cyclists, and pedestrians were evaluated, respectively. In addition, the corresponding selected sites were T-intersection, intersection, and nonsingle intersection. The results of object loss rate were shown in Tables XIV–XVI.

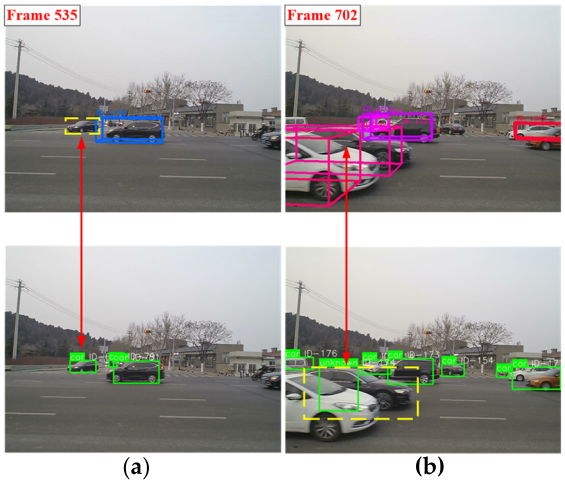

Fig. 17. Schematic of object loss for the vehicle. (a) Object can be detected in 2-D but lost in 3-D. (b) Object can be detected in 3-D but lost in 2-D.

As shown in the tables above, the proposed method can effectively alleviate the problem of object loss. As for various road users, the object loss rates of the proposed tracking method are 0%, 6.82%, and 0%, respectively. We can find that, whatever the object is, the proposed method can get the best performance than other methods. As shown in Fig. 17, take the tracking results of the vehicle as an example. From frames 535 to 568, the vehicle with a yellow rectangular box cannot be detected and tracked by the 3-D tracking method, whereas this object can be tracked by the 2-D tracking method. For the 3-D tracking method, a vehicle was lost during this period. However, due to the proposed method including 2-D and 3-D tracking information in the meantime, the information of the object not being tracked by the 3-D tracking method can be provided by the 2-D tracking method. In addition, an object not being captured by the 2-D tracking method, such as the vehicle in frames 702–709, can be tracked by the 3-D tracking method. In conclusion, the proposed method can solve the problem of the lost object for the 2-D or 3-D tracking method.

Acquiring the trajectory as completely as possible is the goal of the tracking method. Due to the severe occlusion or far distance, a complete trajectory of the object can be divided

TABLE XIII
COMPARISON FOR TRACKING RANGE AMONG DIFFERENT TRACKING METHODS

Scene	Object tracking method	The maximum of tracking distance(m)	The view range of object tracking
T-intersection	DeepSort	100	View range: 90°
	AB3DMOT	51.02	Area radius: 51.02m
	Ours	100	View range: 90° + Area radius: 51.02m
The merging area of interchange	DeepSort	110	View range: 90°
	AB3DMOT	75.73	Area radius: 75.73m
	Ours	110	View range: 90° + Area radius: 75.73m
Intersection	DeepSort	120	View range: 90°
	AB3DMOT	57.28	Area radius: 57.28m
	Ours	120	View range: 90° + Area radius: 57.28m
Roundabout	DeepSort	100	View range: 90°
	AB3DMOT	52.15	Area radius: 52.15m
	Ours	100	View range: 90° + Area radius: 52.15m
Non-signal intersection	DeepSort	45.5	View range: 90°
	AB3DMOT	39.13	Area radius: 39.13m
	Ours	45.5	View range: 90° + Area radius: 39.13m

TABLE XIV
RESULTS OF MISSED OBJECT AMONG DIFFERENT TRACKING METHODS (CAR)

Object tracking method	East-west direction			South-north direction		
	The number of Car	The number of lost car	The object loss rate	The number of Car	The number of lost car	The object loss rate
DeepSORT	32	3	9.375%	77	0	0%
AB3DMOT		0	0%	77	7	3.90%
Ours		0	0%	77	0	0%

TABLE XV
RESULTS OF MISSED OBJECT AMONG DIFFERENT TRACKING METHODS (CYCLIST)

Object tracking method	The number of cyclist	The number of lost cyclist	The object loss rate
DeepSORT	44	3	6.82%
AB3DMOT		29	65.91%
Ours		3	6.82%

TABLE XVI
RESULTS OF MISSED OBJECT AMONG DIFFERENT TRACKING METHODS (PEDESTRIAN)

Object tracking method	The number of pedestrian	The number of lost pedestrian	The object loss rate
DeepSORT	20	0	0%
AB3DMOT		6	30%
Ours		0	0%

into many disconnected trajectories. How to link multiple trajectories belonging to one object plays an important role in tracking.

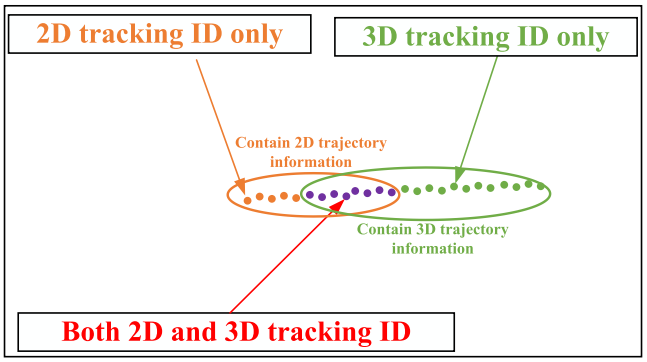

Fig. 18. Schematic of fixing disconnected trajectories.

In general, as shown in Fig. 18, a complete trajectory can be divided into three types of disconnected trajectories. The orange point means that the trajectory only possessed a 2-D tracking ID. The green point indicated the trajectory only had a 3-D tracking ID. As for the purple point, the corresponding trajectory contained both 2-D and 3-D tracking IDs. That is, the trajectory in the orange circular contained 2-D trajectory information, and in the green circular, those points contained 3-D trajectory information. The purple point can be utilized as a bridge to link these three disconnected trajectories as a complete trajectory.

TABLE XVII
RATE OF FIXING DISCONNECTED TRAJECTORIES

The category of object	The number of detected trajectories	The number of trajectories after fixing	Percentage of fixed trajectories with the proposed method
Vehicle	124	108	12.9%
Cyclist	25	15	40%
Pedestrian	34	21	38.24%

Fig. 19. Performance comparison of tracking methods in the dark.

In this article, the repairing rate of disconnected trajectories was utilized to evaluate the performance of linking disconnected trajectories. The elaboration of calculating the repairing rate of disconnected trajectories is given as follows. The number of trajectories was assumed to be *z1* after employing the proposed method to process data collected by roadside multisource platform. According to the 2-D and 3-D tracking IDs between various disconnected trajectories, these trajectories can be linked. After linking these disconnected trajectories to one object, the number of trajectories can de be denoted as *z2*. Then, the definition of the repairing rate of disconnected trajectories was represented by the following equation [8], [34], [35]:

$$F = \frac{z_1 - z_2}{z_1}. \tag{33}$$

The results of fixing the disconnected trajectories among vehicles, cyclists, and pedestrians were shown in Table XVII. We can know that the repairing rates of disconnected trajectories of the proposed method are 12.9%, 40%, and 38.24% for vehicle, cyclist, and pedestrian. The proposed method can effectively link the disconnected trajectories belonging to the same object for various road users. In addition, the proposed method was also tested in the dark. As shown in Fig. 19, due to the headlight, the vehicle with a red rectangular box cannot be detected by the 2-D detection method. However, because the quality of point clouds was unaffected by the light, the vehicle can be detected and tracked by PointRCNN-Selayer and AB3DMOT. Therefore, the proposed method was robust in light.

IV. CONCLUSION

This article developed a novel method to extract road users' trajectories based on the fusion of 2-D and 3-D trajectory information. The proposed method was evaluated at five selected sites. According to the experiment results, the proposed tracking method had a good performance in accuracy and robustness. The proposed object detection method can detect various road users more accurately. By studying the relationship between the number of point clouds and the distance, the weight coefficients α and β were determined. According to the validation experiment, the results showed that the proposed tracking method can significantly alleviate the problem of object loss and fix the disconnected trajectories. Of course, it should be noted that there are still some limitations in this study. The weight coefficients α and β were determined from the perspective of the distance in this article. The other perspectives will be included in the future. The accuracy analysis was just based on the constant speed; the variable speed experiment was not involved. Future studies will focus on these abovementioned issues.

REFERENCES

[1] A. Li, L. Luo, and S. Tang, "Real-time tracking of vehicles with Siamese network and backward prediction," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2020, pp. 1–6.
[2] J. Azimjonov and A. Özmen, "A real-time vehicle detection and a novel vehicle tracking systems for estimating and monitoring traffic flow on highways," *Adv. Eng. Informat.*, vol. 50, Oct. 2021, Art. no. 101393.
[3] C. Lee and J.-H. Moon, "Robust lane detection and tracking for real-time applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 12, pp. 4043–4048, Dec. 2018.
[4] S. Hua, M. Kapoor, and D. C. Anastasiu, "Vehicle tracking and speed estimation from traffic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 153–160.
[5] N. N. S. Hlaing, M. Naing, and S. S. Naing, "GPS and GSM based vehicle tracking system," *Int. J. Trend Sci. Res. Develop.*, vols. 3, nos. 4, pp. 271–275, Jun. 2019.
[6] Z. Tang *et al.*, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8797–8806.
[7] X. Cao, J. Lan, X. R. Li, and Y. Liu, "Automotive radar-based vehicle tracking using data-region association," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8997–9010, Jul. 2022.
[8] X. Song *et al.*, "Augmented multiple vehicles' trajectories extraction under occlusions with roadside LiDAR data," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21921–21930, Oct. 2021.
[9] Y. Fang, H. Zhao, H. Zha, X. Zhao, and W. Yao, "Camera and LiDAR fusion for on-road vehicle tracking with reinforcement learning," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1723–1730.
[10] T. Clunie, M. DeFilippo, M. Sacarny, and P. Robinette, "Development of a perception system for an autonomous surface vehicle using monocular camera, LIDAR, and marine RADAR," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 14112–14119.
[11] W. Song, S. Li, T. Chang, A. Hao, Q. Zhao, and H. Qin, "Cross-view contextual relation transferred network for unsupervised vehicle tracking in drone videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1707–1716.
[12] Z. Liu *et al.*, "Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6640–6653, Jul. 2022.
[13] X. Zhao, P. Sun, Z. Xu, H. Min, and H. Yu, "Fusion of 3D LIDAR and camera data for object detection in autonomous vehicle applications," *IEEE Sensors J.*, vol. 20, no. 9, pp. 4901–4913, May 2020, doi: 10.1109/JSEN.2020.2966034.
[14] Y. Zhang, T. Wang, K. Liu, B. Zhang, and L. Chen, "Recent advances of single-object tracking methods: A brief survey," *Neurocomputing*, vol. 455, pp. 1–11, Sep. 2021.

[15] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, Mar. 2020.

[16] D. Frossard and R. Urtasun, "End-to-end learning of multi-sensor 3D tracking by detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 635–642.

[17] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, "Robust multi-modality multi-object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2365–2374.

[18] J. Luiten, T. Fischer, and B. Leibe, "Track to reconstruct and reconstruct to track," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1803–1810, Apr. 2020.

[19] M. Simon *et al.*, "Complexer-YOLO: Real-time 3D object detection and tracking on semantic point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[20] L. Zhu, X. Geng, Z. Li, and C. Liu, "Improving YOLOv5 with attention mechanism for detecting boulders from planetary images," *Remote Sens.*, vol. 13, no. 18, p. 3776, Sep. 2021.

[21] D. Hu, "An introductory survey on attention mechanisms in NLP problems," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2019, pp. 432–448.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[23] Q. Wang *et al.*, "ECA-Net: Efficient channel attention for deep convolutional neural networks," 2020, *arXiv:1910.03151*.

[24] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[26] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 770–779.

[27] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[28] H. W. Kuhn, "Variants of the Hungarian method for assignment problems," *Nav. Res. Logistics Quart.*, vol. 3, no. 4, pp. 253–258, 1956.

[29] F. Farahi and H. S. Yazdi, "Probabilistic Kalman filter for moving object tracking," *Signal Process., Image Commun.*, vol. 82, Mar. 2020, Art. no. 115751.

[30] T. Kim and T.-H. Park, "Extended Kalman filter (EKF) design for vehicle position tracking using reliability function of radar and LiDAR," *Sensors*, vol. 20, no. 15, p. 4126, Jul. 2020.

[31] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[32] X. Weng, J. Wang, D. Held, and K. Kitani, "3D multi-object tracking: A baseline and new evaluation metrics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10359–10366.

[33] J. Wu, H. Xu, Y. Sun, J. Zheng, and R. Yue, "Automatic background filtering method for roadside LiDAR data," *Transp. Res. Rec.*, vol. 2672, no. 45, pp. 106–114, 2018.

[34] Y. Cui, H. Xu, J. Wu, Y. Sun, and J. Zhao, "Automatic vehicle tracking with roadside LiDAR data for the connected-vehicles system," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 44–51, May/Jun. 2019, doi: 10.1109/MIS.2019.2918115.

[35] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors," *Transp. Res. C, Emerg. Technol.*, vol. 100, pp. 68–87, Mar. 2019, doi: 10.1016/j.trc.2019.01.007.

Shujian Wang received the B.Eng. degree from Qingdao Technological University, Qingdao, China, in 2004. He is currently pursuing the Ph.D. degree with the School of Civil Engineering, Shandong University, Jinan, China.

He is a Vice General Manager with Shandong Hi-Speed Construction Management Group Company Ltd., Jinan. His interests include intelligent transportation systems, project management, and traffic safety.

Rendong Pi received the B.Eng. degree from the School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, China, in 2018, and the M.Eng. degree from the School of Qilu Transportation, Shandong University, Jinan, China, in 2022.

He is currently a Research Assistant with the Autonomous Systems Laboratory, The Hong Kong Polytechnic University, Hong Kong. His current interests include autonomous driving, mobile robots, and intelligent transportation systems.

Jian Li received the M.Eng. degree from Chiba University, Chiba, Japan, in 2011.

He is currently the Deputy Director of the Science and Technology Innovation and Development Department, Shandong Hi-Speed Construction Management Group Company Ltd., Jinan, China. His interests include intelligent transportation systems.

Xinming Guo received the B.Eng. degree from Shanghai Normal University, Shanghai, China, in 2001. She is currently pursuing the M.Eng. degree with Shandong University, Jinan, China.

Her interests include intelligent transportation systems, geographic information systems, and traffic safety.

Youfu Lu is currently a Senior Engineer with Shandong Hi-Speed Construction Management Group Company Ltd., Jinan, China.

Tao Li received the B.Eng. degree from the China University of Petroleum (East China), Qingdao, China, in 2011, and the M.Eng. degree from the School of Civil Engineering, Shandong University, Jinan, China, in 2017.

He is currently the Director of the Department of Traffic Engineering, Shandong Hi-Speed Construction Management Group Company Ltd., Jinan. His current interests include intelligent transportation systems.

Yuan Tian received the B.Eng. and M.Eng. degrees from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2014 and 2017, respectively, and the Ph.D. degree from the School of Engineering, University of Nevada, Reno, NV, USA, in 2021.

He is currently a Lecturer with Shandong University. His interests include intelligent transportation systems, geographic information systems, and traffic safety.