Inter-Vehicle Distance Estimation Method Based on Monocular Vision Using 3D Detection

Ting Zhe, Liqin Huang, *Member, IEEE*, Qiang Wu, *Senior Member, IEEE*, Jianjia Zhang, Chenhao Pei, and Liangyu Li

***Abstract*—Most autonomous vehicles build their perception systems on expensive sensors, such as LIDAR, RADAR, and high-precision Global Positioning System (GPS). However, cameras can provide richer sensing at a considerably lower cost, this makes them a more appealing alternative. A driving assistance system (DAS) based on monocular vision has gradually become a research hotspot, and inter-vehicle distance estimation based on monocular vision is an important technology in DAS. There are still constrains in the existing methods for estimating the inter-vehicle distance based on monocular vision, such as low accuracy when distance is larger, unstable accuracy for different types vehicles, and significantly poor performance on distance estimation for severely occluded vehicles. To improve the accuracy and robustness of ranging results, this study proposes a monocular vision end-to-end inter-vehicle distance estimation method based on 3D detection. The actual area of the rare view of the vehicle and the corresponding projection area in the image are obtained by 3D detection method. An area-distance geometric model is then established on the basis of the camera projection principle to recover distance. Our method shows its potential in complex traffic scenarios by testing the test set data provided on the real-world computer vision benchmark, KITTI. The experimental results have superior performance than the existing published methods. Moreover, the accuracy of occluded vehicle ranging results can reach approximately 98%, while the accuracy deviation between vehicles with different visual angles is less than 2%.**

***Index Terms*—Monocular vision, inter-vehicle distance estimation, area-distance geometric model, 3D detection.**

I. Introduction

VEHICLES have evolved into systems that integrate mechanical, chemical, electronic, computer, and other engineering technologies to ultimately enable self-driving vehicles that exhibit artificial intelligence. However, autonomous vehicles still need to overcome several technical obstacles, such as object recognition, vehicle detection, inter-vehicle distance estimation, and collision warning. Therefore, a driver assistance system (DAS) has developed into an important research field. DAS plays an important role in preventing rear-end collision, reducing traffic accidents, and improving driving safety.

In the context of DAS, inter-vehicle distance estimation is a crucial component of numerous safety-critical applications. Vehicle distance estimation methods include sensor-based methods, vision-based methods, and a combination of multiple sensors, such as the method which combines RADAR and vision sensor [8], [9], [10], or combines a camera and a laser range finder [42]. The latter utilizes the advantages of each independent system to achieve precise ranging. Sensor-based system mainly uses RADAR, LIDAR, and other active sensors [12] to detect the surrounding environment. This system can provide accurate distance information of the target vehicle. However, the high cost of these sensors and the difficulty of collecting data regarding the target vehicles are still critical issues to be solved. In general, vision-based system can be roughly divided into stereo vision and monocular vision. Stereo vision [18], [19] uses stereo image pairs for stereo matching and generates disparity maps, and then obtains the depth information of the target vehicle, thus, we can recover the distance information of the target vehicle. It$'s$ more intuitively and accurately for the long-distance target vehicles to calculate. However, stereo vision systems exhibit low efficiency, require a long execution time and considerable computational complexity due to the complexities of calibrating, and matching between two cameras. Monocular vision [20], [21] does not have the above problems of stereo vision. Because monocular vision has the advantages of low cost, simple structure and wide application range, it is more suitable for embedding into DAS than other systems. Monocular vision-assisted driving system [22] combines the advantages of visual system and assisted driving system, which not only effectively control the real-time performance, but also can be applied to the driving scene of modern vehicles to ensure the safety of driving. Therefore, one of the important components of monocular vision-DAS, namely, inter-vehicle distance estimation based on monocular vision, has gradually become a direction of attention for more and more researchers. However, existing inter-vehicle distance estimation methods based on monocular vision still suffer from problems, such as low precision, large errors in long-distance measurement results, and a narrow the scope of the application scene.

Recently, many monocular vision based methods for inter-vehicle distance estimation have been proposed. Experimental

Manuscript received August 2, 2019; revised November 25, 2019 and January 11, 2020; accepted February 19, 2020. Date of publication March 2, 2020; date of current version May 14, 2020. This work was supported by the Major Science and Technology Projects in Fujian, China, under Grant 2018H0018. The review of this article was coordinated by Dr. A. Chatterjee. *(Corresponding author: Liqin Huang.)*

Ting Zhe, Liqin Huang, Jianjia Zhang, Chenhao Pei, and Liangyu Li are with the College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China (e-mail: n171127046@fzu.edu.cn; hlq@fzu.edu.cn; zhangkienka@gmail.com; n161120081@fzu.edu.cn; n171120065@fzu.edu.cn).

Qiang Wu is with the School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: qiang.wu@uts.edu.au).

Digital Object Identifier 10.1109/TVT.2020.2977623

(a) (b)

Fig. 1. (a) 2D and (b) 3D detection bounding boxes.

results have indicated that monocular vision distance estimation methods can be roughly divided into relative depth estimation and absolute distance estimation. The relative depth estimation method [23], [24] mainly outputs a depth map to represent the depth change of the whole image, and uses different gray values to represent the depth of each pixel in the image. Finally, we can extract the distance information of the target vehicles. However, This method of obtaining distance information can only get the relative relationship between the target vehicle and camera, and the absolute distance of the target vehicle in meters in the traffic scene cannot be obtained. Moreover, the redundant information is contained in the depth map (such as the sky, street lights, road signs, and roadside trees) in traffic scenarios, and thus the efficiency of estimating the target vehicle$'s$ distance is reduced. Article [25] presents an absolute distance estimation method in which an absolute distance value of vehicles in front can be obtained. However, the estimating absolute distance estimation methods based on monocular vision all use the vehicle detection method to extract the required vehicle projection information and then estimate the absolute distance value. Existing object detection based on absolute distance estimation can be divided into 2D detection based and 3D detection based. Methods based on 2D detection [1], [4], [28] apply 2D detection technology [26], [27], [29] to detect an object in an image and locate the object candidate region, the information of objects is represented by 2D bounding box or 2D mask, by using these 2D boxs or 2D masks we can extract required information (e.g., vehicle width, height, projected area, and position or a key point in the image) and complete the distance estimation of the vehicle. Thus, the performance of a 2D detector is the primary factor that affects final ranging accuracy.

Many high-precision 2D detection methods have been proposed. To improve detection efficiency and real-time performance, [45] proposals a real-time vehicle detection algorithm which fuses vision and Lidar point cloud information. At present, the detection effect of popular high-performance 2D object detectors [14], [16] indicates the position and contour shape of the target vehicle in the image through a 2D bounding box. Although recent research has shown that the 2D detection of a large number of object categories exhibits a good performance, using only 2D bounding boxes to recognize objects is insufficient in many practical applications, the intelligent driving ranging system is one example. A 2D bounding box contains considerable redundant information, which cannot reflect the original 3D effect of a vehicle, as shown in Fig. 1(a). Thus, it cannot obtain detailed valuable information for the target vehicle. When estimating distance using a vehicle$'s$ projected information, the physical definition of the projected information is inconsistent with that of the actual information. Consequently, an error occurs in the estimated distance and ranging accuracy is reduced. In article [3], [4], a projection geometric model is proposed to estimate the target vehicle distance by using the projected area obtained by the instance segmentation method [29]. Compared with previous methods, the vehicle can obtain a more detailed projected contour and the redundancy is reduced. However, the 2D mask obtained using this method can neither segment the contour shape of each part of the vehicle in detail nor reflect the stereo pose of the vehicle in the actual scene. When the distance estimation model is established using the projected area, confirming whether the selected vehicle$'s$ projected information exhibits a corresponding projection relationship with actual vehicle information is impossible. Therefore, using 2D projected information in distance estimation will produce a considerable absolute error. However, vehicles are rigid objects with distinctive parts. If we want to accurately estimate the distance value of the vehicle, it is very important to obtain the three-dimensional information of vehicle.

Recently, several methods [41], [47] propose to estimate the vehicle pose by calculating 6 Degree of Freedom (DoF) information to obtain vehicle$'s$ 3D information. Another related approach [7], [13] have explored 3D object bounding box detection for driving scenarios and are most closely related to our proposed method. 3D object detection methods can recover both the 6 DoF pose and the dimensions of an object from an image. Currently, the 3D detection method based on monocular vision [44] has achieved good results, as shown in Fig. 1(b). This method is able to produce 3D bounding box, which can clearly separate the different parts of vehicle and solve the problems of 2D detector. However, a method for estimating inter-vehicle distance based on 3D detection has not been proposed yet. The existing 3D detection method [6] can obtain the 3D information of a vehicle (e.g., size and center point coordinates) and calculate the absolute distance between the target vehicle and the camera by using such information. Nevertheless, this ranging method still suffers from problems, such as large errors in the estimation results between vehicles with different distances and visual angles, low accuracy of long-distance estimation results, and weak robustness of the distance estimation system.

Although many papers have proposed distance estimation methods based on monocular vision, these methods still have several main problems, that can be summarized as follows: 1) The corresponding projection relationship between the real information and the projection information of the vehicle is not taken into account, thus, the inaccurate projection information leads to the error of the ranging result. 2) Because the distance estimation method based on 2D detection does not clearly segment each part of the vehicle, the incorrect vehicle information input in the geometric model of distance estimation leads to the inaccurate of the distance. 3) The distance estimation method based on 3D detection does not fully utilize the camera projection principle, result in the mathematical method for estimating the distance is not high in applicability.

To solve these problems, we improved the method for calculating inter-vehicle distance in [6], and used the advantages of the Deep Convolutional Neural Network and the KITTI dataset [11] to train the 3D detection network. The actual dimension of the vehicle and its projection 3D bounding box can be obtained by 3D detection network; therefore, the actual information and the projection information of the vehicle can be obtained. In order to estimate the absolute distance of the vehicle accurately, we established an area-distance geometric model following the principle of camera projection. The extracted vehicle information above is used as the input of the model, finally, the distance value of the vehicle is obtained through the ranging model.

In summary, to solve the above mentioned problems, we propose an improved distance estimation method. The main contributions of this paper are summarized as follows:

1) In this paper, each part of the projected vehicle in the image is clearly segmented by using the advantages of 3D detection, so that the needed vehicle projection information can be extracted accurately, and the projection relationship between the vehicle information obtained in the image and the actual vehicle information can be guaranteed to be consistent.
2) Based on the principle of camera projection, we establish an area-distance geometric model, which uses the projection relationship between the projected back area of the vehicle and the real back area of the vehicle. Based on this model, a mathematical formula for distance estimation of various types of vehicles is derived.
3) In order to improve the efficiency of distance estimation system, we propose an end-to-end distance estimation framework based on monocular vision by taking advantage of deep learning network and end-to-end framework.
4) In order to evaluate the performance of the ranging method in more detail, we further divide the test set into three test subsets. Three different subsets are used to verify the distance estimation system from different aspects in this paper. The experimental results show that the ranging error in different distance ranges are obviously reduced, and the accuracy of occluded vehicle ranging results can reach approximately 98%, while the accuracy deviation between vehicles with different visual angles is less than 2%.

The remainder of this paper is organized as follows. Section II reviews the state-of-the-art and related works done in field of distance estimation research so far. Section III mainly introduces the overall design idea of the ranging method and the realization method of each independent module in the distance estimation system. Section IV explains the experimental environment and test results, which show the accuracy and robustness of the overall system and the reliability of independent modules. Finally, Section V presents the conclusions and future works.

II. Related Work

In recent years, several methods have been developed to estimate the absolute distance of vehicles based on monocular vision. Geometric models established by distance estimation systems can be classified into three types. The first type involves deriving a model based on a geometric relationship. In [1], [30], the geometric positional relationship of a vehicle in the camera model was used to derive the correspondence between the key points in the image coordinate system and the world coordinate system, and a ranging model was established to realize vehicle ranging. This type of method requires accurate measurement of a camera$'s$ azimuth and elevation angles. Otherwise, the accuracy of distance measurement would considerably be reduced and accurately measuring the elevation angle of a camera on a moving vehicle is difficult. The second type is based on the mathematical regression modeling method. For example, article [31] utilized the correspondence between different standard distances and their positions in an image to build the regression model and measure distance. [46] also uses the idea of fitting regression. This paper combines the vehicle image information acquired by camera to train a distance regression model to achieve the distance estimation task. However, this kind of methods require to collect a considerable amount of training data, then analyze and construct the regression model. It will increases the complexity of the distance estimation and reduces its efficiency. The third type of methods are based on camera imaging model. In [25], [43], distance was measured using the width of a vehicle. These methods are merely suitable for the case where moving vehicles are in well front. When the target vehicle is not exactly in front, projected vehicle width part and the actual vehicle width is not consistent. If this problem is ignored, the robustness and accuracy of the distance estimation will decrease.

According to the different projection principles, monocular visual ranging methods can be divided into two categories: distance estimation based on inverse perspective mapping (IPM) principle and distance estimation based on perspective projection principle. The first type of method [1], [31] use the IPM projection principle to restore the target vehicle$'s$ the absolute distance value. First, the original image is converted into a bird$'s$ eye view (aerial view), and the road surface information of overlooking angle is restored. Then, the converted IPM image is used to estimation the target vehicle distance. However, This kind of method has two shortcomings: 1) The required brightness is relatively high, because when the luminance of the acquired image is low, the performance of the detection system will be reduced and the accuracy of the distance estimation will be reduced. 2) When the original image is converted into IPM image, its size will change, so that the information of some target vehicles in IPM image will be lost, which limits the ranging range of the system.

To avoid the missing information problem in the process of image conversion, a geometric model based on the perspective projection principle was proposed to estimate the absolute distance of the target vehicle. Bao *et al.* [20] presented a linear relationship model based on the vehicle$'s$ average width and the vehicle$'s$ ground true distance to achieve monocular vision ranging. Huang *et al.* [32] proposed a method to measure the longitudinal distance of the target vehicle based on the camera projection principle and the position relationship between the vehicle and the vanishing point in the image. A distance estimation method based on the projection principle is proposed [33], which is realized by using the vehicle width information obtained by

vehicle detection method, and the method also takes into account two kinds of road environments: with and without lane markings. However, the aforementioned methods represent the position and shape of the object vehicle in the image in the form of a 2D box. This representation method cannot obtain more details of the vehicle, and will contain a lot of redundant information. Huang *et al.* [3], [4] proposed a method for calculating the object vehicle$'s$ the distance value based on the projected area obtained via vehicle instance segmentation. Compared with the measurement methods that simply using 2D box information (i.e., vehicle width, height, and position), their methods greatly reduce redundant information and improve ranging accuracy.

Although the 2D mask based on instance segmentation can obtain the detailed projection contour of a vehicle, the overall ranging idea is still inter-vehicle distance estimation based on 2D detection information, this results in several problems, such as the contour shape of each part of a vehicle and the stereo pose of a vehicle not being described as well as the relationship between vehicle projected and actual information not being clearly presented. Therefore, directly using 2D projection information for distance estimation leads to a considerable error.

The development of modern object detection methods has been well-advanced. Numerous methods based on monocular visual object detection have been proposed in recent years. These methods can be easily divided into two categories according to different visualization effects, namely, 2D [17], [29] and 3D [6], [7] detection based. A 2D detection method does not provide all the information required in an actual environment and only presents the position, size, and classification of an object in the image. Although the detection results of existing 2D detection methods are excellent [40], they cannot reflect the 3D pose of an object. Moreover, segmenting the contour shape of each part of an object is impossible. In the real world, objects have 3D shapes. Most practical application scenarios require 3D information, such as the length, width, height, and direction angle of an object. In the application scenario of automatic driving, obtaining the 3D shape and information of a vehicle in an image is crucial. Therefore, 3D detection is currently a research hotspot. The end-to-end model architecture is a key point of a vehicle distance estimation model [15]. In recent years, developments in deep learning techniques and relevant end-to-end technologies have been extensively applied to autonomous driving [38], [39]. To solve the shortcomings of inter-vehicle distance estimation methods based on 2D detection by utilizing the advantages of 3D detection and end-to-end framework, we propose an area-distance end-to-end ranging geometry framework based on the camera projection principle to estimate inter-vehicle distance.

III. System Model

A. System Overview

The primary reason for existing problems in 2D detection ranging methods is that the 3D stereo pose of a vehicle is not taken into account. In actual traffic scenarios, the target vehicle is stereo. If it is regularized, then it can be represented by a real 3D rectangular box, and the parts of the vehicle will be clearly divided. Thus, the projection component of the vehicle in different driving states should exhibit a one-to-one projection relationship with the actual scene. However, as shown in Fig. 1(a), the vehicle projection information provided by a 2D box does not exhibit a corresponding projection relationship with actual vehicle information. If the projection relationship is considered the same, then the distance estimation using this projection relationship will increase the absolute error of the ranging result.

In addition, a 2D detection method in real traffic scenarios suffers from the problem of low recall rate for severely occluded and long-distance vehicles. Thus, we cannot obtain the projection information of occluded and long-distance vehicles. This condition leads to certain limitations in the range and applicability of the overall ranging system. Such issue is crucial in practical applications. The proposed 3D detection method can solve the problems of 2D detection methods possessed. It not only ensures the accuracy of the detection result and improves the recall rate, but also obtains the 3D stereo representation of a vehicle.

In summary, we use the advantages of 3D detection to establish an area-distance geometric model based on the principle of camera projection to recover inter-vehicle absolute distance. The system framework is illustrated in Fig. 2. First, the RGB image is the input of the trained vehicle 3D detection network. Then, we can obtain the actual dimension of the vehicle and the vehicle projected 3D bounding box through the 3D detection network. Subsequently, we use the width and height of the actual dimension of the vehicle and the projected 3D box to calculate the actual and projected areas of the rear part of the vehicle. Finally, in accordance with the principle of camera projection, the area-distance geometric model is established using the area projection relationship to estimate the distance between vehicles. The proposed overall distance estimation system is a complete end-to-end ranging framework that combines object classification, object 3D box, and object absolute distance value. It not only accelerates ranging speed, but also improves ranging performance and detection results. The 3D candidate region adopts the design concepts and deep network framework of the 3D bounding box estimation in [6]. We use the training set prepared beforehand to train the network. Through the trained network, the stable 3D properties of the vehicle can be obtained. Then, the 3D box of the vehicle in the image is visualized to identify the 3D candidate area of the vehicle.

The following sections mainly introduce the design of the partial module of the distance estimation system, which primarily includes the estimation of the vehicle$'s$ physical dimension and the projected 3D bounding box, the actual and projected areas of the rear part of the vehicle, and the design of the distance estimation module.

B. Estimation of the Vehicle$'s$ Physical Dimension and the Projected 3D Box

In general, a 3D stereo box is the smallest box that surrounds the target object in a real 3D world, 3D detection algorithm can generate the projected 3D box of the target vehicles in the image, and it is a corresponding projection relation with the

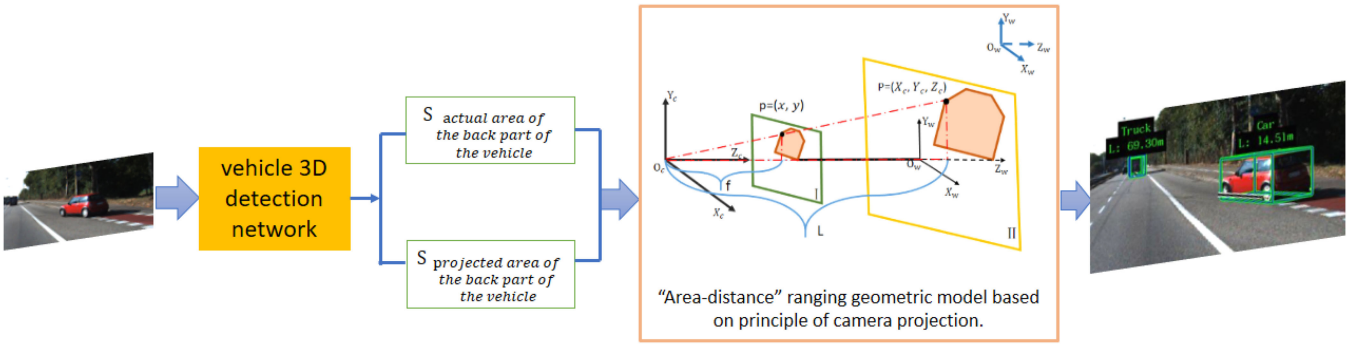


Fig. 2. Framework of the distance estimation system.

real 3D stereo box of the target vehicles. Most target vehicles in the automatic driving scenario are placed horizontally on the ground. Assume that a vehicle is driving on horizontal ground. Then, the elevation φ and roll α angles can be set to zero relative to the horizontal plane. We establish a ranging model under the camera coordinate system. The coordinates of the 3D bounding box vertices can be described by the vehicle$'s$ dimension. Accordingly, the distance value can be calculated in accordance with the conversion relationship between the object and camera coordinate systems and the projection principle. Therefore, obtaining the physical dimension of the target vehicle through 3D detection provides an important foundation for the subsequent ranging work. To obtain the physical dimension of the vehicle, we adopt the principle of bounding box regression in the Faster R-CNN [14] network and the design idea of the dimension estimation architecture in [6]. Based on the convolution feature map of the last layer, we modified the regression parameters after the fully connected layers (FC) and then used the KITTI detection dataset to train the dimension estimation module we needed. Finally, the vehicle dimension information is obtained.

To better estimate the projected 3D box of the vehicle in the image, the experience and advantages of existing 2D target detection methods [34] are fully utilized. The 3D box of the vehicle projection in the image is estimated in accordance with the principle of perspective projection and the geometric constraints posed by the fact that the vehicle 3D box and 2D detection windows fit closely in terms visual appearance.

C. Extraction of Projected and Actual Areas at the Back of the Vehicle

Existing ranging studies indicate that a geometric model established on the basis of the projection relationship of the back area of a vehicle can solve the problem of an unclear projection relationship, and avoid the bottleneck encountered when using other vehicle information for ranging. For example, the method for measuring distance using the vehicle width projection relationship is more suitable for the front vehicle. Because the vehicle$'s$ width profile projected by non-front vehicles may be tilted and twisted. If calculated in accordance with the original projection relation, then the overall ranging accuracy will be reduced. If this type of vehicle is disregarded, then the application scope of the ranging system will be narrowed. Similarly, the same problem appears when the height projection relationship is used. A bottleneck is created when selecting the height value because the left and right height values of the rectangular box of a non-front vehicle may differ. In summary, we use the projection relationship of the back area of a vehicle to establish a ranging model that can be applied to vehicles with different observation angles.

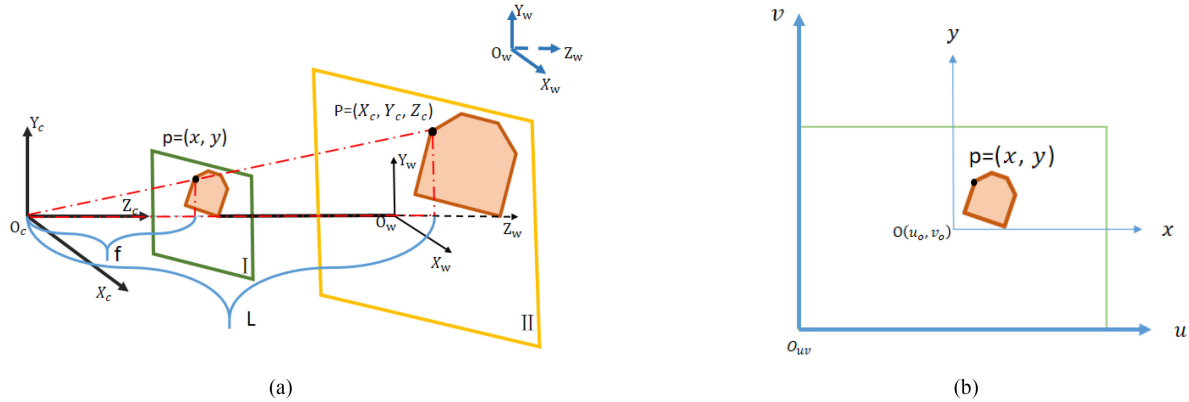
In Section III-B, we have obtained the physical dimensions (length, width, and height) of a vehicle in the image and its projected 3D box. On this basis, we can further calculate the projected (S_{vbp}) and actual (S_{vba}) areas of the rear part of the vehicle, where $S_{vba} = height \times width$.

D. Distance Estimation Module Design

According to the principle of camera projection and the area projection relation of the target vehicle, we first use the projection area of the rear part of the vehicle, the actual area of the rear part of the vehicle, and the focal length of the camera (in pixels) to establish the projection geometry model of distance estimation, as shown in Fig. 3(a), Then the mathematical formula for distance estimation of the target vehicle is derived by using this mathematical geometric model, which is suitable for various types of vehicles. The detailed explanation can be seen in Section I, II, and III.

Compared with the method in [33], vehicle information is more comprehensively utilized to improve the accuracy and applicability of the distance estimation system. Compared with the method in [6], the reliability of the ranging model and the logical rigor of the ranging formula are enhanced to ensure the accuracy of the calculation results. Moreover, we focus on the conversion relationship between the actual back of the vehicle and the back of the projection, and the projection relationship is more clearly defined than that in [3], [4]. Thus, the accuracy of ranging can be improved.

1) Principle of Camera Projection: The principle of camera projection is a method to transform three-dimensional coordinates into two-dimensional coordinates. In order to obtain the form of pixels in the image, we need to transform the four



(a) (b)

Fig. 3. Projection geometry model of distance estimation, where f is the camera focal length and L is the physical distance between the object vehicle and the camera, the orange irregular figure is the rear part of the target vehicle. (a) Principle of camera projection. (b) Image plane to pixel plane.

coordinate systems. First, the points (X_w, Y_w, Z_w) in the world coordinate system is converted to the points (X_c, Y_c, Z_c) in the camera coordinate system, which then become points (x, y) on the 2D plane through perspective projection. Lastly, points (x, y) are stored in the form of pixels (u, v).

a) The conversion of the world coordinate system to the camera coordinate system: If the world coordinate system is in the position shown in Fig. 3(a), then R = I (unit matrix), T = $[0\ 0\ L]^T$, $Z_w = 0$, and Equation (1) can be obtained.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & L \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{bmatrix} \tag{1}$$

b) The conversion of camera coordinate system to image coordinate system: The camera projection principle is applied as shown in Equation (2).

$$Z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{2}$$

c) The conversion of image coordinate system to pixel coordinate system: As shown in Fig. 3(b), the orange irregular figure is the projected part of the target vehicle. The image coordinate system is converted to the pixel coordinate system, as shown in Equation (3).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3}$$

d) The conversion of world point to pixel point: Next, in Equation (4), the conversion relationship between the actual and pixel points in the camera coordinate system can be obtained.

$$\begin{aligned} Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} &= L \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & L \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} f_x & 0 & u_0 & u_0 L \\ 0 & f_y & v_0 & v_0 L \\ 0 & 0 & 1 & L \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ 0 \\ 1 \end{bmatrix}, \end{aligned} \tag{4}$$

where $\frac{f}{d_x} = f_x$, $\frac{f}{d_y} = f_y$, $(u_0, v_0) = (0, 0)$, and $Z_c = L$, Equation (4) is transformed into Equation (5), and then we can get the projection relationship between the points of the target vehicle, such as p and P points in Fig. 3(a).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{L} \begin{bmatrix} f_x X_c + u_0 L \\ f_y Y_c + v_0 L \\ L \end{bmatrix} = \frac{1}{L} \begin{bmatrix} f_x X_c \\ f_y Y_c \\ L \end{bmatrix} \tag{5}$$

2) Relationship of Area Conversion Is Derived From the Relationship of Point Conversion: The actual area of the rear part of the target vehicle is divided into N parts along the Y_c direction, and each part is approximately a rectangle as shown in Fig. 4. The four vertices of the i-th rectangle are marked as P_1^i, P_1^{i+1}, P_2^i, and P_2^{i+1}, where $P_r^i = (P_{rx}^i, P_{ry}^i) = ({x_r}^i, {y_r}^i)$,

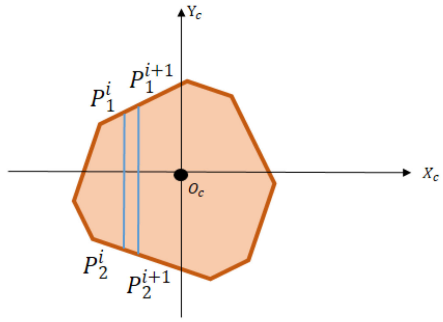


Fig. 4. Actual area of the rear part of the target vehicle.

$(r = 1, 2; i = 1, 2, 3, \ldots, N)$. P_{rx}^i and P_{ry}^i represent the X_c and Y_c coordinates of the four coordinate points, respectively.

Then, the actual area of the rear part of the target vehicle can be expressed as follows:

$$\begin{aligned} S &= \sum_{i=1}^{N} \left(P_{1y}^i - P_{2y}^i\right) \left(P_{2x}^{i+1} - P_{2x}^i\right) \\ &= \sum_{i=1}^{N} \left(y_1^i - y_2^i\right) \left(x_2^{i+1} - x_2^i\right). \end{aligned} \tag{6}$$

Using the projection relationship between the key points of the target vehicle, i.e. Equation (5), we can further deduce the projection relationship between the back area of the target vehicle, and the following equation can be obtained:

$$\begin{aligned} S &= \left[\sum_{i=1}^{N} \left(v_1^i - v_2^i\right) \left(u_2^{i+1} - u_2^i\right)\right] \frac{L^2}{f_x f_y} \\ &= S_{projection} \frac{L^2}{f_x f_y}, \end{aligned} \tag{7}$$

where $S_{projection}$ represents the projected area of the target vehicle in an image, and S represents the actual area of the vehicle.

3) Estimating the Physical Distance of the Target Vehicle: Based on Equation (7), a mathematical representation of the distance estimation is proposed for the target vehicle, i.e., Equation (8) as follows:

$$L = \left(\frac{f_x f_y S}{S_{projection}}\right)^{\frac{1}{2}} = \left(\frac{f_x f_y S_{vba}}{S_{vbp}}\right)^{\frac{1}{2}}, \tag{8}$$

Where L is the physical distance of the vehicle in front, $f_x = f_y = 7.2153 \times 10^2$, S_{vba} is the actual area of the rear part of the vehicle, and S_{vbp} is the projected area of the rear part of the vehicle.

IV. Experiment

The proposed distance estimation system is mostly applied to the modern vehicle automatic driving system in the actual traffic scene. we use a camera mounted behind the windshield of a vehicle for capturing images.

Datasets: The ranging method proposed in this study primarily involves the vehicle$'s$ 3D detection network, which outputs the 3D information parameters of the vehicle, and can be used to visualize the 3D box projected by the vehicle in the image. In the proposed distance estimation system, the involved network model requires 3D information of the vehicle (i.e., length, width, and height) during training. Thus, the dataset used must contain the label of the real 3D information of the vehicle. Nowadays, the computer vision evaluation dataset which is mostly used for vehicle image analysis which contains the largest automatic driving scene in the world is the KITTI dataset [11]. This dataset provides the corresponding real 3D annotation information for each moving object in a camera$'s$ field of view. Therefore, we mostly use the object detection data in KITTI to train the network and then test and verify it on the KITTI detection benchmark. At present, other large datasets based on vehicle monocular vision lack the ground truth of a vehicle$'s$ 3D information and the vehicle$'s$ distance required by our ranging system. Therefore, we used the KITTI dataset to train and test the proposed distance estimation system.

Our distance estimation system is mainly designed for the assistance driving system of modern vehicles, thus, we only focus on the "vehicle" category. KITTI contains a training set of 7481 images and a test set 7518 images. However, there are no ground truth label in the KITTI test dataset, so we separated a part of the data from the KITTI training set as the test set of the experiment according to the pre-set rules. The pre-set rules are as follows: First, the data in the training and test sets must come from different video sequences. Second, the selected test data should satisfy the following conditions: different distance ranges, occlusion degrees, and visual angles. According to this rule, we used 3981 images from the training set as the test set to verify and analyze our distance estimation method.

To present the performance changes of the ranging method in detail, we divided the test set into three test subsets.

- Test subset for different vehicle distances (including similar visual angles and occlusion degrees). Vehicle samples with different actual distances are mostly extracted to form this subset, which is used to verify the accuracy of our method in different distance ranges.
- Test subset with different occlusion degrees (including similar visual angles and vehicle distances). The occlusion rate of a vehicle in the KITTI test set is divided into three levels: 0 (visible), 1 (partly occluded), and 2 (fully occluded). We selected vehicles with occlusion degrees greater than 1, extracted them as test samples, and formed test subsets with different occlusion degrees to verify the ranging effect of our distance estimation system for vehicles with different occlusion degrees.
- Test subset with different visual angles (including similar vehicle distances and occlusion degrees). We determined a vehicle$'s$ position by the definitions provided in the international vehicle collision warning system [36]. Vehicles are divided into two types: front and sideway vehicles. A front vehicle implies no deviation between the longitudinal centerlines of the subject and target vehicles. If a deviation exists, then the target vehicle is a sideway vehicle. From these definitions, we can organize the test subsets of different visual angles and use the data from this subset to

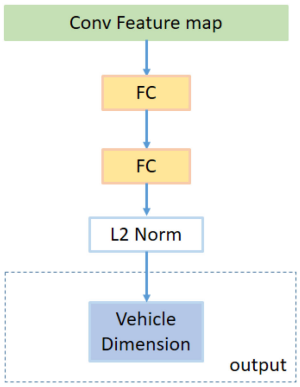


Fig. 5. Dimension estimation module.

verify the variation of the ranging accuracy of our distance estimation system for front or sideway vehicles.

Dimension estimation module and 3D box estimation: The entire vehicle 3D detection network is implemented based on the CNN network framework. The dimension information of the vehicle can be output through the network, and the projected 3D box of the vehicle in an image can be visualized using the dimension information in the camera coordinate system. Therefore, obtaining the dimension information of the vehicle is crucial. We train a deep CNN network, modify the parameter regression part of the network, and then use the network to obtain the dimension parameters of the vehicle. To obtain the vehicle parameters we need, during the training, each input image is resized to 224×224, and then add our dimension estimation module to the pre-trained VGG network [35] without the FC layers. In the KITTI dataset, there are many different categories, such as vehicles, trucks, buses, etc. The distribution of object dimensions for category instances is low-variance and unimodal, for example, the dimension variance for different types of vehicles are approximately several centimeters. Therefore, we used L_2 loss directly. The dimension estimation module is shown in Fig. 5.

Evaluation metrics:

a) Absolute Error: $\Delta(m) = |L_{ground\ truth} - L_{experimental}|$.

b) Relative Error: $\delta(\%) = |\frac{\Delta}{L_{ground\ truth}}|$.

c) Average Error: $\overline{\Delta}(m) = \frac{1}{n}\sum_{i=1}^{n}|\Delta_i|$.

b) Average Error Rate: $\overline{\delta}(\%) = \frac{1}{n}\sum_{i=1}^{n}|\frac{\Delta_i}{L_{ground\ truth\ i}}|$.

To verify the accuracy and robustness of the proposed distance estimation system, we tested and analyzed the ranging method developed in this study from the following aspects: the verification of the entire system model, the verification of the effectiveness of the independent module on the entire ranging accuracy, and the visualization of the ranging results.

A. Verification of the Entire Distance Estimation System Model

1) Accuracy Changes at Different Distances: In the test subset of different vehicle distances, we divided the distance range into three main categories: 0–10 m, 10–20 m, and >20 m. Then, we used our ranging method to compare the average error of the ranging results in different distance ranges with existing ranging methods [1], [2], [4] to verify the accuracy and robustness of the overall ranging system. The results are provided in Table I.

TABLE I
AVERAGE ERROR OF DIFFERENT DISTANCE ESTIMATION METHODS ($meters = m$)

Distance range	*0-10m*	*10-20m*	*>20m*
method [1]	0.74	1.77	6.52
method [2]	0.81	1.81	7.21
method [4]	0.38	0.85	2.17
ours	**0.164**	**0.327**	**0.396**

Methods [1], [2], and [4] use 2D detection to achieve distance estimation. The projection relationship between vehicle projection information obtained via 2D detection (e.g., vehicle width and height) and the real vehicle information exhibit a deviation, because the 2D detection results cannot reflect the stereoscopic pose of the vehicle result in the average error relatively high.

Our 3D bounding box based on 3D detection is formed on the basis of the real 3D stereo box projection of a vehicle, so the projection information of each part of a vehicle (e.g., width, height, and back area) corresponds to real vehicle information. The detection result of 3D detection is more stereoscopic and accurate than that of 2D detection, vehicle pose can be expressed in much more details, this enables us to extract accurate back projection information regarding the vehicle and estimate the vehicle$'s$ distance value.

The experimental results show that our method is best for ranging in different distance ranges. Especially, the average error of our ranging results is guaranteed to be less than 0.5 m in distances longer than 20 m. Moreover, the maximum deviation between the average errors is approximately 0.3 m, and the error between different distance ranges is significantly reduced. Thus, the entire distance estimation system is more stable and robust.

2) Accuracy Changes Under Different Occlusion Degrees: In the test subset of different occlusion degrees that we collected, these occluded vehicles mainly exhibit partial or severe occlusion. Then, we used our method to estimate the absolute distance of two occluded vehicles separately and compared the result with ground truth to verify the applicability and accuracy of the distance estimation system to occluded vehicles. The experimental results are presented in Table II.

The test results indicate that the proposed ranging method can be used to estimate the distance of occluded vehicles. Under different occlusion conditions, the average error can be controlled within 0.5 m for an occluded vehicle within 25 m. The accuracy of the ranging result can reach 98%, and the effect remains considerable.

3) Accuracy Changes Under Different Visual Angles: To evaluate the performance of our method in detail and verify the robustness and applicability of the system, we established test subsets with different observational visual angles. Both front and sideway vehicles are tested on the test subset, and the average error rate of the estimated distance value is compared with the methods presented in [3], [4], [37]. The results are provided in Table III.

TABLE II
DISTANCE VERIFICATION OF OCCLUDED VEHICLES

Number [1]	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Occlusion degrees	1	2	2	1	1	1	2	2
Ground truth (m)	9.045	9.665	12.395	13.19	16.125	16.315	22.3	24.095
Experimental (m)	8.983	9.74	12.281	13.259	15.936	16.518	22.524	24.368
Absolute Error (m)	0.062	0.075	0.114	0.069	0.189	0.203	0.224	0.273
Relative Error $(\%)$	0.685	0.776	0.920	0.523	1.172	1.244	1.004	1.133
Occlusion category	Partly Occluded				Fully Occluded			
Average Error (m)	0.366				0.389			
Accuracy $(\%)$	**98.38**				**97.94**			
Overall Average Error (m)	***0.377***							
Overall accuracy $(\%)$	***98***							

[1]The "Number" represents the index number of the vehicle samples.

TABLE III
COMPARISON BETWEEN EXPERIMENTAL DISTANCE AND GROUND TRUTH OF TWO GROUPS OF EXPERIMENTS

	Number [1]	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Distance verification of front vehicles	Ground truth(m)	5.37	7.835	12.61	14.52	20.03	24.775	36.29
	Experimental distance(m)	5.37	7.79	12.46	14.62	19.924	24.488	36.538
	Absolute Error(m)	***0***	***0.045***	***0.15***	***0.1***	***0.106***	***0.287***	***0.248***
Distance verification of sideway vehicles	Ground truth(m)	6.79	7.78	10.49	13.34	15.63	20.29	34.7
	Experimental distance(m)	6.761	7.737	10.304	13.184	15.463	20.534	35.056
	Absolute Error (m)	***0.029***	***0.043***	***0.186***	***0.156***	***0.167***	***0.244***	***0.356***

Average error rate of vehicle distance estimation from different observation visual angles

Method	Front vehicle$'s$ average error rate$(\%)$	Sideway vehicle$'s$ average error rate $(\%)$
Method[37]	4.702	9.237
Method[3]	1.333	4.698
Method[4]	1.077	2.820
Ours	**0.370**	**1.750**

[1]The "Number" represents the index number of the vehicle samples.

The experimental results indicate that our method is optimal for vehicle ranging results with different visual angles compared with other methods. The average error rate of the front vehicle ranging result is less than 0.5%, and the sideway vehicle ranging result is reduced to approximately 1.75%. Compared with other methods, a significant drop for error is observed. This drop is sufficient to show that adding a 3D detection module is advantageous for the distance estimation system. In addition, the error rate between the sideway and front vehicles is less than 2%. Compared with those of other methods, the deviation of the ranging accuracy between vehicles with different viewing angles is significantly reduced. Consequently, the limitations and inapplicability of existing ranging methods are addressed and the robustness of the system is enhanced.

B. Independent Module Verification

1) Verifying the Impact of the 3D Detection Module on the Accuracy of the Entire Ranging System: The primary concept of our ranging method is to use the information of the projected area at the back of a vehicle to achieve distance measurement. Therefore, the method of obtaining the projected area on the back of the vehicle is the key to determine the accuracy of the distance estimation results. We compared the absolute error values of

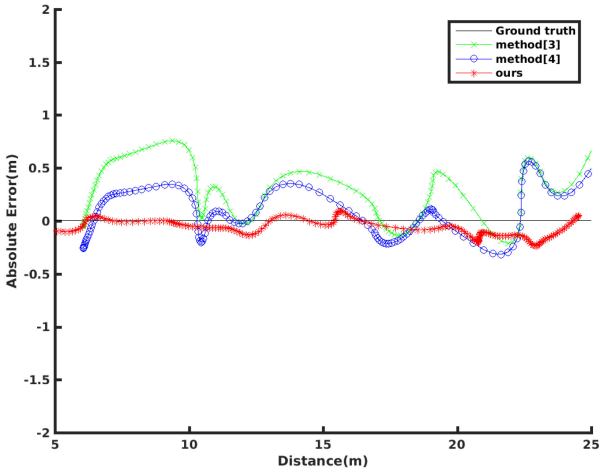


Fig. 6. Absolute error values for different distance ranges.

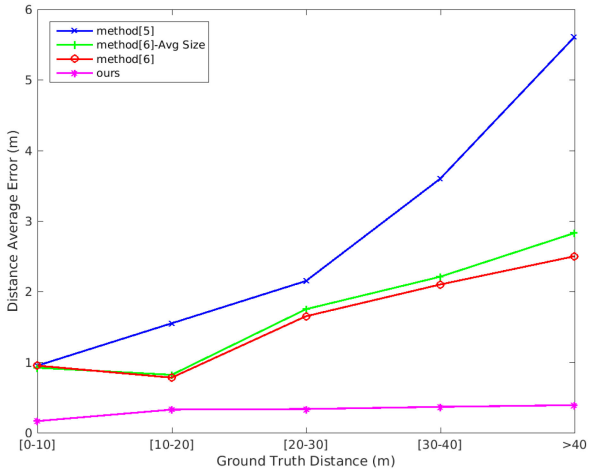


Fig. 7. Average error value over different distance ranges.

different distance ranges of our method with an existing method that uses the area-distance geometric model but without adding a 3D detection module for ranging [3], [4]. The verification results are presented in Fig. 6.

The method presented in [3], [4] obtains the projected area at the back of a vehicle on the basis of instance segmentation and then uses the area projection relationship to realize distance estimation. In the existing method [3], the mask of entire vehicle is directly used as the projection mask around the back of the vehicle. Ideally, the distance between different parts of vehicle and the camera should be different. However, the existing mask covers all parts of vehicle, so the estimated distance is not precise. In order to improve the accuracy, we need to segment the vehicle parts before projection. The method developed in [4] estimates the projected area of the back of a vehicle using the geometric relationship of the entire mask and the change in the attitude angle of the vehicle. However, the projected portion of the back of the vehicle segmented from the entire mask exhibits an irregular 2D shape. Determining whether the segmented projected back portion of a vehicle corresponds to the actual back of the vehicle is difficult because no clear 3D stereo shape exists. Thus, the accuracy of ranging is low.

We can obtain the stereo shape of a vehicle in an image through 3D detection, and all parts of the vehicle are clearly segmented. Thus, the projection relationship between the projected and actual parts is more consistent. Consequently, the fuzzy problem of the aforementioned area projection relation is solved. The experimental results show that the absolute error of the ranging result of our method is the smallest among different distance ranges, and the disparity with ground truth has the lowest fluctuation. In particular, when the distance is longer than 20 m, the error of the ranging result is significantly reduced compared with the methods proposed in [3], [4]. Therefore, the concept of using 3D detection module to extract the information on the back of the vehicle to realize distance estimation is effective and accurate.

2) Verifying the Impact of the Area-Distance Geometry Model on the Accuracy of the Entire Ranging System: Our method is compared with the 3D detection module but not with area projection relation modeling [5], [6]. The accuracy comparison diagram of the ranging results is provided in Fig. 7.

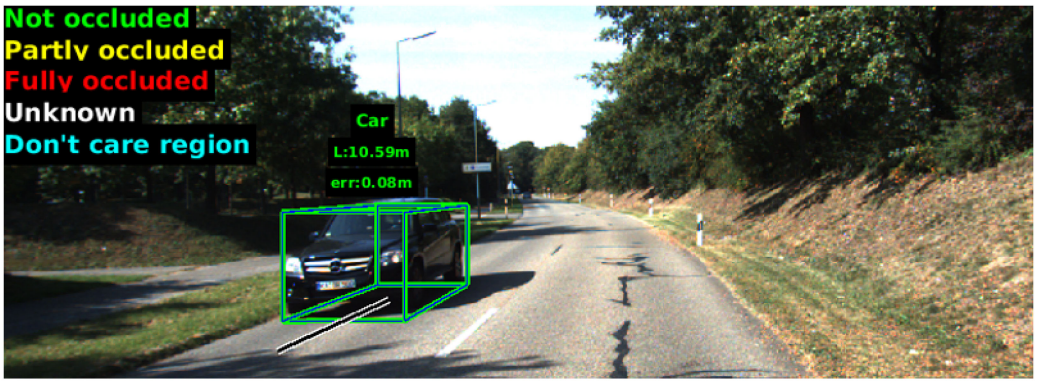
Methods developed in [5], [6] use a vehicle$'s$ physical dimension and view angle information to calculate position information, and the obtained position information is calculated on the basis of the vehicle$'s$ center point Z coordinate and the dimension. In an actual traffic scenario, however, the driving direction of a vehicle is uncertain, and the observed visual angle of the object vehicle is different. This method is unsuitable for estimating the absolute distance of all the vehicles. The use of this method will lead to lower overall ranging accuracy, a narrower application range, and less robustness of the distance estimation system. The geometric model of area-distance is established by using the projection relationship between the actual area and the projected area on the back of the vehicle, deduce the ranging formula, and realize distance estimation. Our method can ensure the reliability of the distance calculation formula.

The comparison in Fig. 7 shows that the proformance the proposed of method is superior, and the average error at different distance ranges is significantly reduced. Compared with the ranging results of methods [5] and [6], especially when the distance of ground truth is more than 40 m, the average error of our distance estimation method is significantly mitigated. Moreover, the accuracy deviation of the ranging results among different distances is the smallest. This finding shows that our distance estimation geometric model is reliable and effective and the concept of overall ranging is logically and mathematically rigorous. It is also applicable to vehicles with various driving directions, thus, the robustness of the distance method is enhanced.

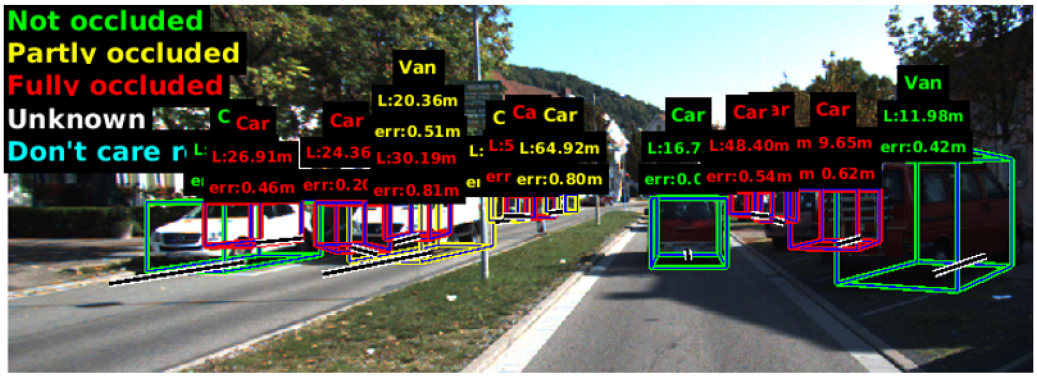
C. Visualized Results

To intuitively demonstrate the advantages of our ranging method, we use ground truth to verify the accuracy of the estimation results. The results are visualized in Fig. 8.

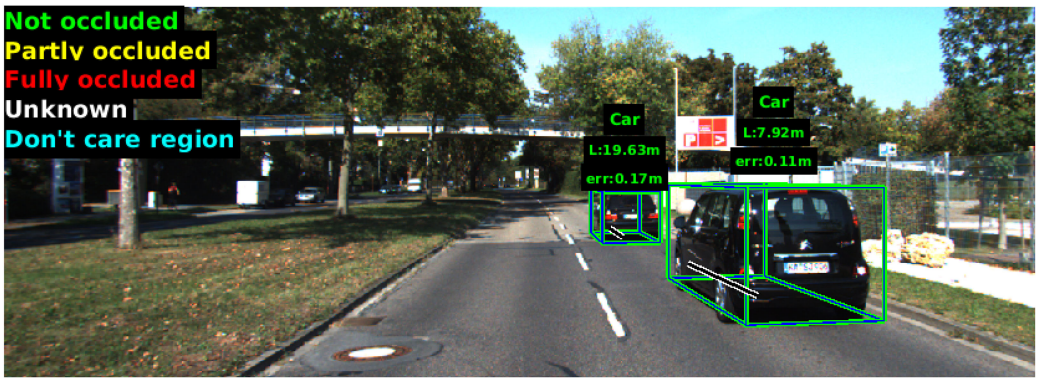
Compared with the scenario shown in the existing methods, this work makes a bit more complicated cases by increasing the



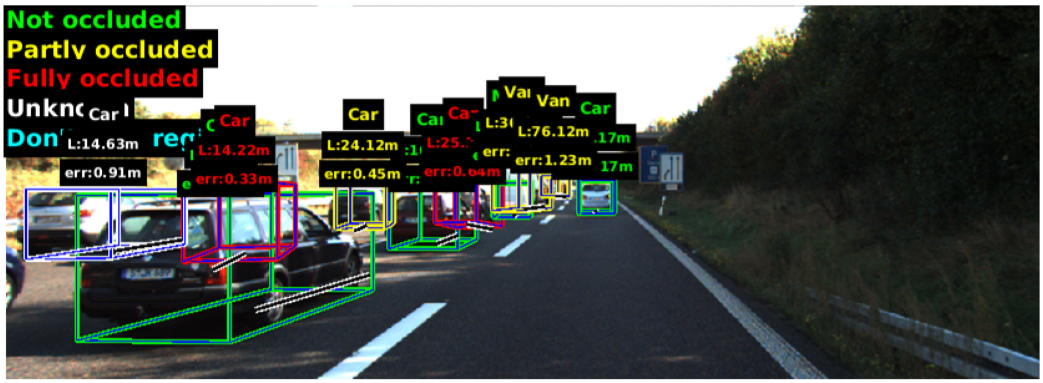
(a)



(b)



(c)



(d)

Fig. 8. Distance prediction on the KITTI dataset. Image (a) and Image (c): Simple traffic scenarios. Image (b) and Image (d): Complex traffic scenarios.

number of vehicles in the scene, changing view angles, using diverse occlusion situations. The proposed distance estimation method is tested based on such complex scenario. Fig. 8 shows the visualization results of the proposed methods under various cases.

According to the test results, it is feasible to build the geometric model of ranging based on the 3D vehicle detection module, and the experimental effect also remains highly impressive. The proposed method not only realizes the stereoscopic detection effect of the target vehicle but also ensures the accuracy of the estimated distance value between vehicles. For a target vehicle within a distance of 25 m, the absolute error of the ranging result is less than 0.5 m. Meanwhile, a target vehicle within the range of more than 60 m has an absolute error of approximately 0.5 m. Even the distance estimation of a severely occluded target vehicle can guarantee the accuracy of the estimation result.

V. Conclusion

This study sets up an area-distance ranging geometry model based on the principle of camera projection. Then, using the advantage of 3D detection, the vehicle 3D detection and the ranging geometry model are combined to propose a robust inter-vehicle distance estimation method based on vehicle-mounted monocular vision. The actual area of the rear part of the target vehicle and the corresponding projected area in the image are obtained using the vehicle 3D detection method. Then, the absolute distance between the vehicles is recovered using the ranging geometry model. To comprehensively verify the performance of the system, we establish test subsets from three different angles and consider accuracy changes at different vehicle distances, occlusion degrees, and observation visual angles. The experimental results show that the proposed method not only adapt to a variety of complex traffic scenarios, but also exhibits high precision and strong robust performance for vehicles under different driving conditions.

In the future work, we will pay attention to more vehicle traffic driving scenarios (e.g., highways, street corners, and rural streets) to further expand the application scope of the distance estimation system. In addition, we will focus on real-time performance to continuously improve our distance estimation system.

References

[1] L. Liu, C. Fang, and S. Chen, "A novel distance estimation method leading a forward collision avoidance assist system for vehicles on highways," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 937–949, Apr. 2017.

[2] S. Sivaraman and M. M. Trivedi, "Integrated lane and vehicle detection, localization, and tracking: A synergistic approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 906–917, Mar. 2013.

[3] L. Huang, Y. Chen, Z. Fan, and Z. Chen, "Measuring the absolute distance of a front vehicle from an in-car camera based on monocular vision and instance segmentation," *J. Electron. Imag.*, vol. 27, no. 4, pp. 1–10, Jul. 2018.

[4] L. Huang, T. Zhe, J. Wu, Q. Wu, C. Pei, and D. Chen, "Robust inter-vehicle distance estimation method based on monocular vision," *IEEE Access*, vol. 7, pp. 46059–46070, 2019.

[5] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, Santa Rosa, CA, USA, 2017, pp. 924–933.

[6] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7074–7082.

[7] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3 D object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vision Patter Recognit.*, 2016, pp. 2147–2156.

[8] Distronic PlusWith Steering Assist, 2013. [Online]. Available: https://www.youtube.com/watch?v=lpIddjyHHVs

[9] F. Garcia, P. Cerri, A. Broggi, A. Escalera, and J. M. Armingo, "Data fusion for overtaking vehicle detection based on radar and optical flow," in *Proc. IEEE Intell. Veh. Symp.*, 2012, pp. 494–499.

[10] A. Haselhoff, A. Kummert, and G. Schneider, "Radar-vision fusion for vehicle detection by means of improved Haar-like feature and AdaBoost approach," in *Proc. Eur. Signal Process. Conf.*, 2017, pp. 2070–2074.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 3354–3361.

[12] R. C. Daniels, E. R. Yeh, and R. W. Heath, "Forward collision vehicular radar with IEEE 802.11: Feasibility demonstration through measurements," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1404–1416, Feb. 2018.

[13] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D voxel patterns for object category recognition," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1903–1911.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[15] H. M. Seliem, R. Shahidi, M. H. Ahmed, and M. Shehata, "On the end-to-end delay in a one-way VANET," *IEEE Trans. Veh. Technol*, vol. 68, no. 9, pp. 8336–8346, Sep. 2019.

[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, Dec. 2015, pp. 1440–1448.

[17] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.

[18] V. T. B. Tram and M. Yoo, "Vehicle-to-vehicle distance estimation using a low-resolution camera based on visible light communications," *IEEE Access*, vol. 6, pp. 4521–4527, 2018.

[19] V. D. Nguyen, T. T. Nguyen, D. D. Nguyen, and J. W. Jeon, "Toward real time vehicle detection using stereo vision and an evolutionary algorithm," in *Proc. IEEE 75th Veh. Technol. Conf.*, 2012, pp. 1–5.

[20] D. Bao and P. Wang, "Vehicle distance detection based on monocular vision," in *Proc. IEEE Int. Conf. Prog. Informat. Comput.*, 2016, pp. 187–191.

[21] A. A Ali and H. A. Hussein, "Distance estimation and vehicle position detection based on monocular camera," in *Proc. Al-Sadeq Int. Conf. Multidisciplinary IT Commun. Sci. Appl.*, 2016, pp. 1–4.

[22] E. Raphael, R. Kiefer, P. Reisman, and G. Hayon, "Development of a camera-based forward collision alert system," *SAE Int. J. Passenger Cars-Mech. Syst.*, vol. 4, pp. 467–478, 2011.

[23] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2650–2658.

[24] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[25] G. Kim and J. S. Cho, "Vision-based vehicle detection and inter-vehicle distance estimation," in *Proc. IEEE Int.Conf. Control, Autom., Syst.*, Jeju Island, South Korea, Oct. 2012, pp. 17–21.

[26] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Kauai, HI, USA, 2001, vol. 1, pp. 511–518.

[27] X. Wen *et al.*, "Improved Haar wavelet feature extraction approaches for vehicle detection," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2007, pp. 1050–1053.

[28] M. Rezaei, M. Terauchi, and R. Klette, "Robust vehicle detection and distance estimation under challenging lighting conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2723–2743, Oct. 2015.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2961–2969.

[30] C. H. Chen, T. Y. Chen, D. Y. Huang, and K. W. Feng, "Front vehicle detection and distance estimation using single-lens video camera," in *Proc. 3rd Int. Conf. Robot, Vision, Signal Process.*, Nov. 2015, pp. 14–17.

[31] P. Wongsaree, S. Sinchai, P. Wardkein, and J. Koseeyaporn, "Distance detection technique using enhancing inverse perspective mapping," in *Proc. 3rd Int. Conf. Comput. Commun. Syst.*, Apr. 2018, pp. 217–221.

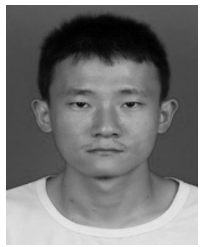
[32] D.-Y. Huang, C.-H. Chen, T.-Y. Chen, W.-C. Hu, and K.-W. Feng, "Vehicle detection and inter-vehicle distance estimation using single-lens video camera on urban/suburb roads," *J. Visual Commun. Image Representation*, vol. 46, pp. 250–259, Jul. 2017.
[33] J. Han, O. Heo, M. Park, S. Kee, and M. Sunwoo, "Vehicle distance estimation using a mono-camera for FCW/AEB systems," *Int. J. Automot. Technol.*, vol. 17, no. 3, pp. 483–491, Jun. 2016.
[34] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 354–370.
[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[36] "Intelligent transport systems-forward vehicle collision warning systems-performance requirements and test procedures," ISO, Geneva, Switzerland, ISO 15623/TC 204, Intelligent transport systems, 2013.
[37] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 740–756.
[38] Q. Wang, L. Chen, B. Tian, W. Tian, L. Li, and D. Cao, "End-to-end autonomous driving: An angle branched network approach," in *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 11599–11610, Dec. 2019.
[39] M. Bojarski *et al.*, "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.
[40] S. H. Qi, J. Li, Z. P. Sun, J. T. Zhang, and Y. Sun, "Distance estimation of monocular based on vehicle pose information," *J. Phys., Conf. Ser.*, vol. 1168, 2019, Art. no. 032040.
[41] D. Wu, Z. Zhuang, C. Xiang, W. Zou, and X. Li, "6D-VNet: End-to-end 6DoF vehicle pose estimation from monocular RGB images," *IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/papers/WAD/Wu_6D-VNet_End-To-End_6-DoF_Vehicle_Pose_Estimation_From_Monocular_RGB_Images_CVPRW_2019_paper.pdf
[42] Z. Liu, D. Lu, W. Qian, K. Ren, J. Zhang, and L. Xu, "Vision-based inter-vehicle distance estimation for driver alarm system," *IET Intell. Transport Syst.*, vol. 13, no. 6, pp. 927–923, Jun. 2019.
[43] G. P. Stein, A. D. Ferencz, and O. Avni, "Estimating distance to an object using a sequence of images recorded by a monocular camera," U.S. Patent 8,164,628 B2, Apr. 24, 2012.
[44] P. L. Liu, "Monocular 3D object detection in autonomous driving—A review," Nov. 26, 2019. [Online]. Available: https://towardsdatascience.com/monocular-3d-object-detection-in-autonomous-driving-2476a3c7f57e
[45] H. Wang, X. Lou, Y. Cai, Y. Li, and L. Chen, "Real-time vehicle detection algorithm based on vision and lidar point cloud fusion," *J. Sensors*, vol. 2019,, 2019, Art. no. 8473980.
[46] F. Gökçe, G. Üçoluk, E. Şahin, and S. Kalkan, "Vision-based detection and distance estimation of micro unmanned aerial vehicles," *Sensors*, vol. 15, no. 9, pp. 23805–23846, Sep. 2015.
[47] J. Yuan *et al.*, "Estimation of vehicle pose and position with monocular camera at urban road intersections," *J. Comput. Sci. Technol.*, vol. 32, pp. 1150–1161, Dec. 2017.

Ting Zhe received the B.S. degree in electronic information science and technology from Shenyang Ligong University, Shenyang, China, in 2017. She is currently working toward the M.S. degree in electronics and communications engineering with Fuzhou University, Fuzhou, China. Her current research interests include image processing, computer vision, and deep learning.

Liqin Huang (Member, IEEE) received the Ph.D. degree from Fuzhou University, Fuzhou, China. He is currently a Full Professor with the College of Physics and Information Engineering, Fuzhou University. His current research interests include image processing, computer vision, artificial intelligence, traffic scene understanding, and medical image processing.

Qiang Wu (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2004. He is currently an Associate Professor and a Core Member of the Global Big Data Technologies Center, University of Technology Sydney. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. He serves as a Reviewer for several journals including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS. His research has been published in many premier international conferences, including ECCV, CVPR, ICIP, and ICPR, and major international journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Jianjia Zhang received the B.S. degree in electronic information science and technology from Fujian Normal University, Fujian, China, in 2017. He is currently working toward the M.S. degree with Fuzhou University, Fuzhou, China. His research interests include computer vision and deep learning.



Chenhao Pei received the B.S. degree in Internet of Things engineering in 2016 from Fuzhou University, Fujian, China, where he is currently working toward the Ph.D. degree in communication and information system. His research interests include lane detection, deep learning, autonomous driving, and computer vision.

Liangyu Li received the B.S. degree in Internet of Things engineering in 2016 from Fuzhou University, Fujian, China, where he is currently working toward the M.S. degree in communication and information system. His research interests are in machine learning and computer vision.