Vehicle-Related Distance Estimation Using Customized YOLOv7

Xiaoxu Liu and Wei Qi Yan

Auckland University of technology, Auckland 1010 New Zealand

Abstract. With the popularity of autonomous driving, the development of ADAS (Advanced Driver Assistance Systems), especially collision avoidance systems, has become an important branch in the field of deep learning. In the face of complex traffic environments, collision avoidance systems need to detect vehicles quickly and accurately in traffic distance to the vehicle in front. Against this background, in this paper, we aim at investigating how to build a fast and robust model for vehicle distance estimation. The theoretical insights are synthesized in the context of odometry and customized YOLOv7 based on what a conceptual framework is proposed. In this paper, KITTI is employed as the dataset for model training and testing. Being one of the pioneer works on distance estimation based on KITTI, the unique value of this research work lies in the first time using YOLOv7 with attention model as a distance estimation model and getting 4.253 on RMSE.

Keywords: Autonoumous vehicles · YOLOv7 · Vehicle detection · Distance estimation · Scene understanding.

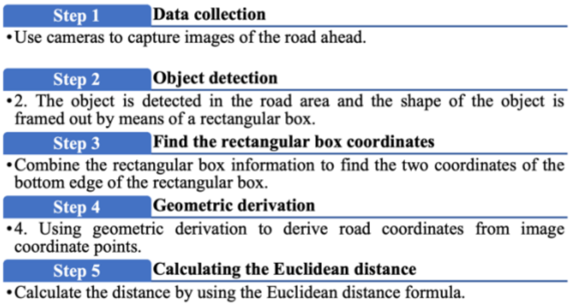
1 Introduction

Advanced Driver Assistance Systems (ADAS) provide a safe and automated driving experience that will reshape our relationship with automobile. In the near future, autonomous vehicles will allow passengers to experience a personalized and interconnected driving experience, given vehicles the ability to sense, act seamlessly and intelligently handle real-time road conditions [36, 37]. Amongst them, vision and radar systems play an important role in ADAS. The vision system is responsible for sensing the surroundings and taking the necessary measures to ensure the safety of all road users. At the same time, the radar systems continuously sense the distance between vehicles in real time, improving driving efficiency and safety [54, 55, 56, 57, 58]. For decades, one of the most popular ideas in the literature for solving distance detection problem pertaining to ADAS is visual object detection of current traffic environment and the distance to surrounding obstacles by means of deep learning methods [27, 28]. An important breakthrough in deep learning-based neural networks is that visual tasks do not have to be coded manually [38, 39, 40, 41]. Deep learning neural networks allow various features to be extracted automatically from training examples [34, 47, 48, 49, 50, 51, 52]. A neural network is considered to have “deep” learning capability if it has input and output layers with at least one implicit intermediate layer [29, 30, 35].

Recent theoretical developments have revealed that the YOLO series is currently one of the most advanced methods for efficient implementation of deep neural networks for vision processing [42, 43, 44]. The YOLO series are much efficient, there is no complex detection process and only the image needs to be fed into the neural network to obtain the detection results. Moreover, YOLO series are very good at avoiding background errors and generating false positives.

For distance detection, a strategy to obtain distance information is through laser detecting and ranging [31, 32, 33]. Laser-based distance measurement is widely considered at the time of developing the Collision Warning System [1,2,3]. However, LiDAR is very complex, expensive and low yielding, which can only be used for testing vehicles at present. In addition, ultrasound, infrared and microwave radar can also be employed for vehicle detection and ranging, but the range of ultrasound and infrared is narrow, microwave radar is susceptible to interference and the reliability of the detection results is weak, while these methods cannot distinguish between detection targets. This poses a few problems while carrying out that the hardware devices such as radar and infrared are expensive and complex to integrate with the camera and have limitations in terms of measurement accuracy. Moreover, few studies have revealed on abandoning costly distance measuring hardware equipment and inferring the distance information from the detected 2D video frames.

Therefore, deep learning-based detection has a promising application prospect and can be combined with the methods to achieve better results. By calibrating the internal and external parameters of the camera, the distance to the vehicle in front is estimated by using the visual projection model principle and geometric ranging methods to warn of a possible collision.



Step 1 **Data collection**

- Use cameras to capture images of the road ahead.

Step 2 **Object detection**

- 2. The object is detected in the road area and the shape of the object is framed out by means of a rectangular box.

Step 3 **Find the rectangular box coordinates**

- Combine the rectangular box information to find the two coordinates of the bottom edge of the rectangular box.

Step 4 **Geometric derivation**

- 4. Using geometric derivation to derive road coordinates from image coordinate points.

Step 5 **Calculating the Euclidean distance**

- Calculate the distance by using the Euclidean distance formula.

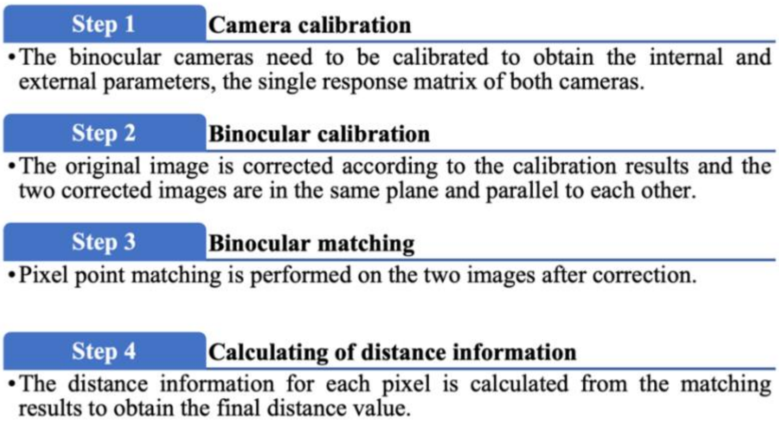
Fig. 1. The steps of monocular ranging method

The existing ranging methods based on visual information comprise two branches: Monocular camera-based ranging methods and binocular camera-based ranging methods. The general principle of monocular camera ranging is to firstly identify the target by image matching and then to estimate the target distance by its size in the image. The general steps shown in Fig.1 are data collection, object detection, finding the rectangular box coordinates, geometric derivation and calculating Euclidean distance. Among them, the circumferential ranging method has a larger fisheye lens distortion, and the circumferential camera is generally employed for low-speed scenes, mainly for detecting ground markings, so the camera lens faces down; the other is the front-view camera ranging, which is characterized by the other is forward-looking camera ranging, which is featured by a smaller aberration of the front-view lens, and the camera is mounted under the rear-view mirror of the car, which can be harnessed in low-speed and high-speed vehicle scenes for detecting vehicles, pedestrians and obstacles in front, so the camera lens have to face forward [4].

Compared to the front view camera ranging method, the circumferential fisheye camera, because the lens faces downwards and the aberration coefficient is large, based on the camera model, the usage of mathematical geometry for ranging is no longer tried and will result in a larger error; the idea is based on the single strain matrix and affine transformation for ranging. The core knowledge is to get the four points of aberration correction map to customize the solution of the single strain matrix corresponding to the four points of the image, which is extremely relevant to the accuracy of the calibration. The front view camera is a normal camera with low aberrations based on the camera model by deriving the relationship between the pixel coordinates and the world coordinates.

Binocular vision imitates human eye structure and takes use of two or more cameras to collect images of different orientations of the same target. The 3D information of target can be accurately calculated by the matching image points between the left and right images under the binocular camera model. Binocular ranging is split into four steps as shown in Fig.2: Camera calibration, binocular calibration, binocular matching, and calculation of the distance.

The advantage of binocular ranging is that there is no limit to the recognition rate, because in principle there is no need for recognition before measurement, but rather all obstacles are measured directly; and there is no need to maintain a sample database for binocular ranging, because there is no concept of a sample for binoculars. The advantages of monocular ranging, on the other hand, are the low cost, the low requirement for computing resources and the relatively simple system architecture.



Step 1 **Camera calibration**

- The binocular cameras need to be calibrated to obtain the internal and external parameters, the single response matrix of both cameras.

Step 2 **Binocular calibration**

- The original image is corrected according to the calibration results and the two corrected images are in the same plane and parallel to each other.

Step 3 **Binocular matching**

- Pixel point matching is performed on the two images after correction.

Step 4 **Calculating of distance information**

- The distance information for each pixel is calculated from the matching results to obtain the final distance value.

Fig. 2. The main steps of binocular method

It is of interest to know whether high precision and high-speed object detection and ranging can be still achieved while keeping costs low. Therefore, the aim of this study

is to develop a more sophisticated model for vehicle detection and ranging using monocular camera. The contributions of this paper are listed as follows:
(1) A proven object detection and ranging model based on monocular camera will be implemented.
(2) To our knowledge, this is the first time using YOLOv7 with attention model as the basic architecture for distance estimation.
(3) A higher detection and ranging performance will be generated.

2 Literature Review

In this section, we review the recent literature on machine vision-based vehicle detection and vehicle ranging. Vehicle detection is the basis for vehicle ranging, while distance estimation from the vehicle ahead provides important data to support vehicle collision avoidance systems. Therefore, more and more computer vision research publications focus on vehicle detection and ranging tasks.

We review two streams of literature: In the first stream of relevant work, we study vehicle detection and ranging with binocular camera. Measuring distance in a real-world coordinate system is the most critical one to be solved for accurate estimation of vehicle speed, a task that is simple when using stereo vision. For each detected vehicle, the relative distance can be obtained directly using the parallax values of the pixels contained in the vehicle area. The basic principle of stereo vision is to observe the same scene from multiple viewpoints (usually two). This enables the obtention of digital images of the three-dimensional scene. By using epipolar geometric principles, three-dimensional shape and position of the surrounding scene is rebuilt [6]. Chui et, al., proposed an algorithm that includes four modules: pre-processing, edge detection, line segment matching, and vehicle search and distance estimation for their multi-resolution stereovision system. The coarse-to-fine detection algorithm is employed for vehicle detection task. In coarse-to-fine detection algorithm, tedious pre-processing operations are inevitable. The pre-processing performs downsampling and low-pass filtering processes that increase computation and slow down detection speed. In the vehicle search and distance estimation module, the front vehicles and their distances will be found and estimated using the average distances of the horizontal and vertical line segments based on the right image [5].

Brojeshwar et al., proposed a more advanced method to deal with the object detection and distance estimation problem. The detection is based on Viola-Jones algorithm where Haar-like features are applied to train a cascaded classifier using the well known AdaBoost algorithm [7, 8, 9]. The distance is obtained by using stereo vision is usage of a novel approach that followed where corners with big eigen values are obtained in segmented regions of both images. Keeping the left image as a reference, for each corer found in the left image, a sub-image is obtained and transferred to the right image using homography and a match is obtained using color correlation. A spiral search is employed for getting the best matching point around the point got by homography. The result is a set of coordinate pairs from both images. A stricter match between this set of points is carried out by using the fact that corresponding points will lie on epipolar lines, and a resultant set of points are arrived at which have exact correspondence in both images. This step was taken to avoid inaccurate matches due to thresholds and differences in the color resolution capabilities of the two cameras [10].

However, the binocular distance estimated by using traditional machine learning methods that require a lot of time for pre-processing and have low accuracy. Moreover, for binocular ranging itself, the computational effort is very high and the performance of the computing unit is very demanding, which makes the productization and miniaturization of binocular systems difficult and costly. Furthermore, the alignment of the binoculars has a direct impact on the accuracy of the distance measurement. The binocular stereo vision method relies on the natural light in the environment to capture images, due to environmental factors such as changes in light angle and light intensity, the brightness of the two images can vary considerably, which can pose a great challenge to the matching algorithm. The binocular ranging method is also not suitable for scenes that are monotonous and lack texture. Since the binocular stereo vision method matches images based on visual features, it can be difficult to match scenes that lack visual features (e.g., sky, white walls, desert, etc.), resulting in large matching errors or even matching failure.

The second stream of literature relevant to our work concerns detection and ranging with a monocular camera. The knowledge of dimensions in the real coordinates of certain features, objects or road sections is a fundamental problem in estimating distances by using monocular systems. This is often referred to as the scale factor for converting pixels to real-world coordinates. Another requirement is to consider the flat road assumption [11, 12, 13]. Firstly, based on indicator lines, augmentation lines or areas [17, 18, 19], these methods do not require calibration of the camera system, but rather measure the actual distance between two or more virtual lines on the road, or the actual size of the road area. The distance estimation is then posed as a detection problem in which all vehicles are detected at the same distance whenever they cross a predefined virtual line or area. Since the virtual line or area is located on the road, an accurate distance estimation involves the precise location of the contact point of a part of the vehicle. This part of the vehicle should be identical at the second position to obtain a consistent estimate of speed [20].

Based on the true size of the object, including the license plate and the vehicle [14, 15, 16, 20], Jong proposed a method of detecting an object and a method of estimating a vehicle's distance from a bird's eye view through inverse perspective mapping (IPM) were applied. In the proposed method, ACFs were employed to generate the AdaBoost-based vehicle detector. The ACFs were extracted from the LUV color, edge gradient, and orientation (histograms of oriented gradients) of the input image. Subsequently, by applying IPM and transforming a 2D input image into 3D by generating an image projected in three dimensions, the distance between the detected vehicle and the autonomous vehicle was detected [21].

Arabi et al., presented a comprehensive solution for distance estimation of the following vehicle solely based on visual data from a low-resolution monocular camera. To this end, a pair of vehicles were instrumented with real-time kinematic (RTK) GPS, and the lead vehicle was equipped with custom devices that recorded video of the following vehicle. Forty trials were recorded with a sedan as the following vehicle, and then the procedure was repeated with a pickup truck in the following position. Vehicle detection and distance estimation were then conducted by employing a DeepStream streaming analytics toolkit and ANN on the video footage [22].

In contrast to the works of Jong [21] and Arabi et al.[22], the work was based on the AdaBoost-based vehicle detector in traditional machine learning, which may yield lower detection accuracy than the models based on advanced deep learning. However, the work takes use of less monetary cost because the model is based on a 2D to 3D transformation to estimate distances rather than relying on a hardware device. In comparison, Arabi costs more money, but invokes more advanced deep learning techniques to detect objects and estimate distance [22].

Following our research and review of the extensive literature, we find that the existing methods on monocular camera-based vehicle detection and ranging are often based on conventional machine learning methods or high monetary cost filming devices, few studies have focused on using deep learning methods to address vehicle detection and ranging at a reduced monetary cost. For the task of estimating distances based on images, deep neural networks have high performance for image processing and modern deep learning frameworks will be by far the best choice for processing image data. Without relying on expensive ranging equipment, deep learning models with improved vehicle detection accuracy can make vehicle distances calculated based on detection frame information more accurate. Therefore, research into vehicle detection and ranging implemented using monocular cameras and deep learning frameworks is urgent and necessary.

3 Our Methods

We are use of YOLOv7 as the underlying architecture for our deep learning models. The purpose of YOLOv7 is to solve two problems. For model structural re-referencing, a planned model structural re-referencing is proposed by using the concept of gradient propagation paths to analyse the structural re-referencing strategies applicable regarding each layer in different networks.

Whilst using a dynamic label assignment strategy, new problems arise in the training of models with multiple output layers, such as how to better assign dynamic targets to the outputs of different branches. To address this issue, the authors proposed a new approach to label assignment called the coarse-to-fine guided label assignment strategy.

YOLOv7 also provided “extend” and “compound scaling” methods for real-time detectors, which allow for a more sophisticated approach. The methods for more efficient use of parameters and computational effort. At the same time, this method can effectively reduce the parameters of a real time detector by up to 50% and offers faster inference and higher detection accuracy [23].

Generally, YOLOv7 firstly resizes the input image to 640×640 and inputs it into the backbone network, then outputs a feature map with three layers of different sizes through the head layer network, and outputs the prediction results through the REP module and the conv module.

Different from the original YOLOv7 model, our proposed model replaces the conv module in backbone with the Convolutional Block Attention Module (CBAM) [24]. The Convolutional Block Attention Module consists of two modules, the Channel Attention (CAM) and the Spatial Attention Module (SAM). The CAM enables the network to focus on the foreground of an image, allowing the network to pay more attention to meaningful regions, while the SAM enables the network to focus on locations in the whole image that are rich in contextual information. The input feature maps were subject to global max pooling and global average pooling based on width and height to obtain two feature maps. Then, they were fed into an MLP with a number of neurons in the first layer and the activation function as ReLU. This two-layer neural network was shared. After that, the MLP output features are subjected to element-wise summation operation and then sigmoid activation operation to generate the final channel attention feature. Finally, the channel attention feature and the input feature map are subject to element-wise multiplication operation and generate the spatial attention module. The function is,

$$M_C(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) = \sigma(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0(F_{max}^c)\right)) \qquad (1)$$

where σ denotes the sigmoid function, $W_0 \in \mathbb{R}^{C/r\times C}$, and $W_1 \in \mathbb{R}^{C\times C/r}$. Note that the MLP weights, W_0 and W_1, are shared for both inputs and the ReLU activation function is followed by W_0.

The feature map output from the Channel attention module is applied as the input feature map for this module. Firstly, we conduct a global max pooling and global average pooling based on the channel and get two feature maps. These two feature maps are then concatenated based on the channel. The spatial attention feature is then generated by sigmoid, and the feature is multiplied by the input feature of the module to obtain the final feature. The function is [45]:

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) = \sigma(f^{7\times7}([F_{avg}^s; F_{max}^s])) \qquad (2)$$

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation with the filter size of 7×7.

After the detection task is completed, we get four numbers in the bounding box, namely (x_0, y_0, *width*, *height*). where x_0 and y_0 are applied to tile or adjust the bounding box. The width and height are adopted for measuring the object and actually describing the detected object and details. The width and height will vary depending on the distance of the object from the camera.

As we know, the image is refracted when passing through the lens because light can also enter the lens and in the case of a mirror, light can be reflected, which is how we get the exact reflection of the image. But in the case of the lens image there is almost no stretching [25].

The monocular camera generates a one-to-one relationship between the object and the image, the relationships of variables are shown in Fig. 3. Using this principle, we can deduce a relationship between known parameters. Using the principle of similar triangles, we can obtain the formulas as follows [25, 26, 46, 53]:

$$\frac{f}{d} = \frac{r}{R} \qquad (2)$$

$$f = d \times \frac{r}{R} \text{ pixels} \tag{3}$$

$$d = f + \frac{R}{r} \text{ cm} \tag{4}$$

where *f* is the focal length, *r* is the height of the measured vehicle in the image, *R* is the height of the vehicle being measured and d is distance from the camera to the object.

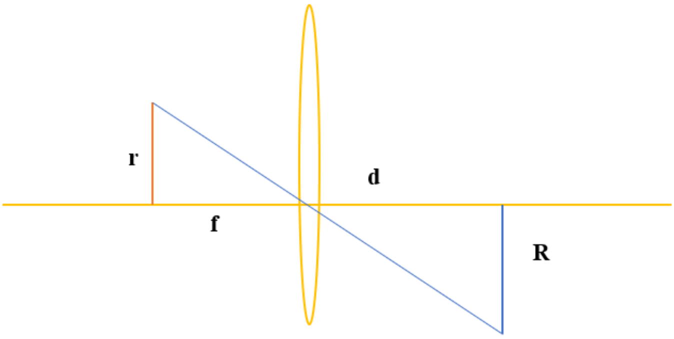


Fig.3. The relationships of variables.

4 Experimental Results

In this paper, we present an advanced deep learning-based vehicle detection and distance estimation model for low-spend monocular cameras. This was experimentally investigated by PYTHON 2.7, RTX5000 GPU and 32GB RAM. Our data samples are from the KITTI dataset. The KITTI dataset contains the internal and external parameters of the in-car camera, as well as the coordinates, width and height of the detection frame. We randomly chose 4,000 samples for developing our deep learning model and divided them into 7:3 for training and test. The outputs demonstrated in this section match state of the art methods. The result in Fig. 4 provides evidence that our modified YOLOv7 with the distance estimation algorithm [25] is able to generate satisfying performance of vehicle detection and distance estimation.

We set appropriate parameter values (*epochs*=3500, *batch_size*=1, *learning_rate*=0.01) to train our modified YOLOv7. The network training process in Fig. 5 shows that losses of both training and validation decrease between 0 epochs and 1,000 epochs, until after 1,000 epochs the loss curves decrease slightly and flattens out around 0.068.

We present a quantitative comparison in the constructed KITTI dataset for all the evaluation metrics in Table. 1. We compare several advanced YOLO models and the transformer model. The results show that YOLOv7 performs significantly better than the other YOLO models and transformer. Furthermore, our modified YOLOv7 that is added the convolutional block attention module performs even better than the original YOLOv7. It may indicate that the delivers significantly better results due to the convolutional block attention module is combined with YOLOv7.

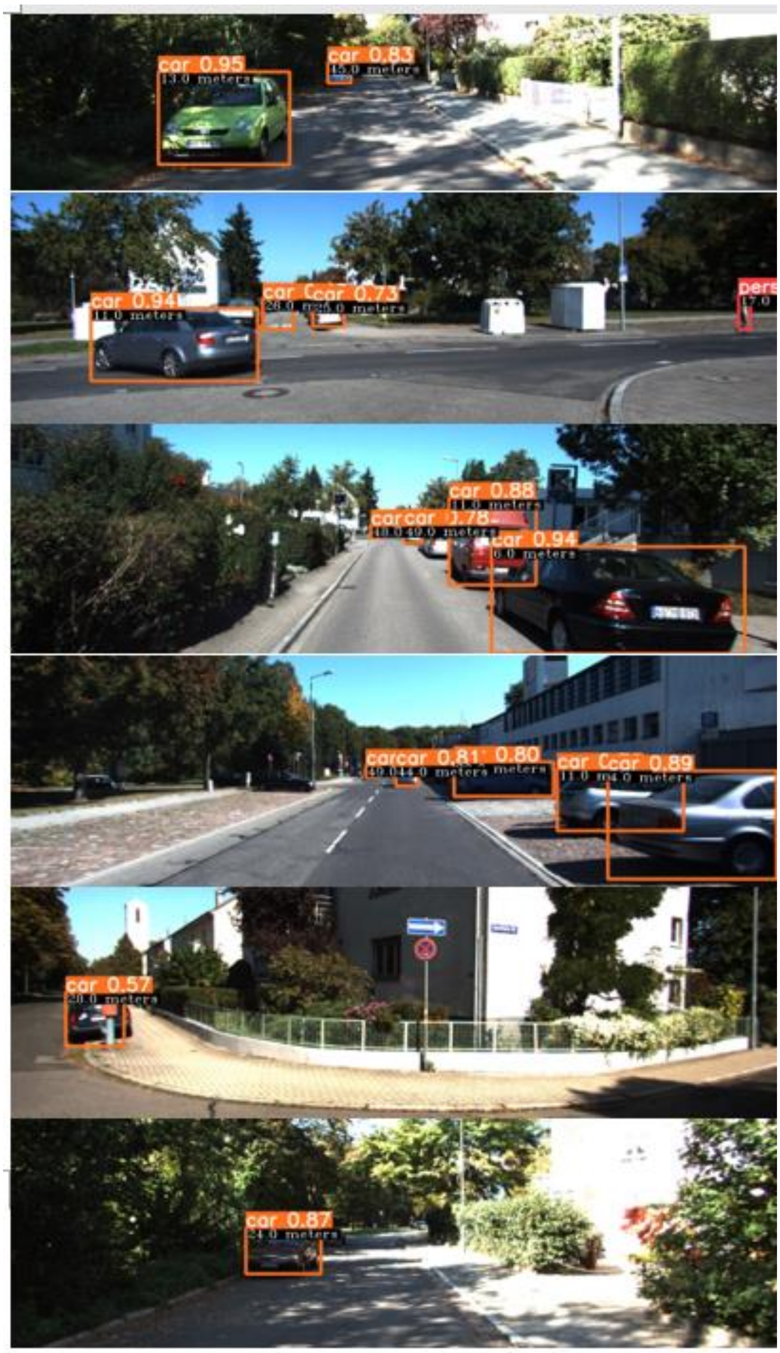


Fig. 4. The example of vehicle detection and distance estimation using the modified YOLOv7

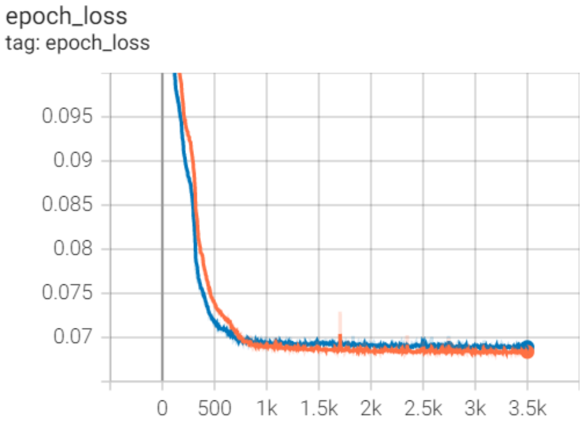


Fig. 5. The training process of the modified YOLOv7. The blue curve indicates the training loss while the orange curve indicates the validation loss.

Table 1. Quantitative comparisons of multiple deep neural networks

Training modality	AbsRel	SqRel	RMSE
YOLOv5	0.182	0.841	4.138
YOLOv6	0.179	0.857	4.516
YOLOv7	0.139	1.651	4.275
Modified YOLOv7	**0.134**	**1.528**	**4.253**
Detection transformer (DETR)	0.211	1.257	4.833

Table 2. Average AbsRel of different neural networks in different distance categories

Training modality	0-10m	10-20m	>20m
YOLOv7	0.121	0.143	0.153
Modified YOLOv7	0.117	0.139	0.173

We also grouped the distances into three categories: 0-10m, 10-20m and >20m. For each category, we calculated the average AbsRel in Table. 2. The results indicate that our modified YOLOv7 outperforms the original YOLOv7 in all the three distance categories. To summarize the findings in Table. 1 and Table. 2, this may raise concerns about object detection and distance estimation tasks that can be successfully addressed by YOLOv7 model and improve performance using the attention module.

5 Conclusion

A combination of YOLOv7 and attention module could enable a low-cost monocular camera-based vehicle detection and distance estimation task with satisfactory results. This project is the first comprehensive investigation of distance estimation by using customized YOLOv7 and the performance of our customized YOLOv7 obtain 4.253 of RMSE. This would be a fruitful area for further work to add other attention mechanisms and receive further enhancements.

References

1. Tinchev, G., Penate-Sanchez, A., & Fallon, M.: Learning to see the wood for the trees: Deep laser localization in urban and natural environments on a CPU. IEEE Robotics and Automation Letters, 4(2), 1327-1334, (2019)
2. Kuznietsov, Y., Stuckler, J., & Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. IEEE Conference on Computer Vision and Pattern Recognition (pp. 6647-6655), (2017)
3. Liao, Y., Huang, L., Wang, Y., Kodagoda, S., Yu, Y., & Liu, Y.: Parse geometry from a line: Monocular depth estimation with partial laser observation. In IEEE International Conference on Robotics and Automation (ICRA) (pp. 5059-5066), (2017)
4. Zhang, J., Hu, S., & Shi, H.: Deep learning based object distance measurement method for binocular stereo vision blind area. International Journal of Advanced Computer Science and Applications, 9(9), (2018)
5. Chiu, C. C., Chung, M. L., & Chen, W. C.: Real-time front vehicle detection algorithm for an asynchronous binocular system. J. Inf. Sci. Eng., 26(3), 735-752, (2010)
6. Zhao, M., Mammeri, A., & Boukerche, A.: Distance measurement system for smart vehicles. International Conference on New Technologies, Mobility and Security (NTMS), pp. 1-5, (2015)
7. Paul, V. and Michael, J., Rapid object detection using a boosted cascade of simple Features. International Conference on Computer Vision and Pattern Recognition (2001)
8. Goncalo, M., Paulo, P. & Urbano, N.,: Vision-based pedestrian detection using HAAR-LIKE features, *Robotica*, 2006.
9. Rainer, L., Alexander, K., & Vadim, P.: An empirical analysis of boosting algorithms for rapid objects with an extended set of Haar-like features. *Intel Technical Report MRL-TR*, 2002.
10. Bhowmick, B., Bhadra, S., & Sinharay, A.: Stereo vision based pedestrians detection and distance measurement for automotive application. International Conference on Intelligent Systems, Modelling and Simulation, pp. 25-29 (2011)
11. Gunawan, A.A.S., et al.: Detection of vehicle position and speed using camera calibration and image projection methods. Procedia Comp. Sci. 157, 255–265 (2019)
12. Kim, J-H., et al.: Reliability verification of vehicle speed estimate method in forensic videos. Foren. Sci. Int. 287, 195–206 (2018)
13. Huang, T.: Traffic speed estimation from surveillance video data. IEEE Comp. Vis. Patt. Rec. (CVPR), pp. 161–165 (2018)
14. Vakili, E., et al.: Single-camera vehicle speed measurement using the geometry of the imaging system. Mult. Tools Apps. 79, 19307–19327 (2020)

15. Llorca, D.F., et al.: Two-camera based accurate vehicle speed measurement using average speed at a fixed point. IEEE Intell. Transp. Sys. Conf. (ITSC), pp. 2533–2538 (2016)
16. Wu, W., et al.: Vehicle speed estimation using a monocular camera. Proceedings of SPIE 9407, Video Surveillance and Transportation Imaging Applications. SPIE (2015)
17. Dahl, M., Javadi, S.: Analytical modeling for a video-based vehicle speed measurement framework. Sensors 20, 160 (2020)
18. Javadi, S., et al.: Vehicle speed measurement model for video-based systems. Comp. & Elec. Eng. 76, 238–248 (2019)
19. Czapla, Z.: Vehicle speed estimation with the use of gradient-based image conversion into binary form. Sig. Proc. Alg. Arch. Arrang. Apps. (SPA), pp. 213–216 (2017)
20. Fernández Llorca, D., Hernández Martínez, A., & García Daza, I.: Vision-based vehicle speed estimation: A survey. IET Intelligent Transport Systems, 15(8), 987-1005 (2021)
21. Kim, J.: Efficient vehicle detection and distance estimation based on aggregated channel features and inverse perspective mapping from a single camera. *Symmetry* 11, no. 10: 1205, (2019)
22. Arabi S., Sharma A., Reyes M., Hamann C., Peek-Asa, C.: Farm vehicle following distance estimation using deep learning and monocular camera images. Sensors, 22(7):2736, (2022)
23. Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. DOI:10.48550/arXiv.2207.02696 (2022)
24. Woo, S., Park, J., Lee, J. Y., & Kweon, I. S.: CBAM: Convolutional block attention module. European Conference on Computer Vision (ECCV) (pp. 3-19).(2019)
25. Khan, M., Paul, P., Rashid, M., Hossain, M., & Ahad, M.: An AI-based visual aid with integrated reading assistant for the completely blind. IEEE Transactions on Human-Machine Systems, pp. 91-99, (2017)
26. Liu, X. Yan, W.: Depth estimation of traffic scenes from image sequence using deep learning PSIVT (2022)
27. Liu, X., Yan, W.: Traffic-light sign recognition using Capsule network. Springer Multimedia Tools and Applications (2021)
28. Liu, X., Yan, W.: Vehicle-related scene segmentation using CapsNets. IEEE IVCNZ (2020)
29. Liu, X., Nguyen, M., Yan, W.: Vehicle-related scene understanding using deep learn. Asian Conference on Pattern Recognition (2019)
30. Liu, X.: Vehicle-related Scene Understanding Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand (2019)
31. Mehtab, S., Yan, W.: FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. ACM ICCCV: (2021)
32. Mehtab, S., Yan, W.: Flexible neural network for fast and accurate road scene perception. Multimedia Tools and Applications (2021)
33. Mehtab, S., Yan, W., & Narayanan, A.: 3D vehicle detection using cheap LiDAR and camera sensors. IEEE IVCNZ (2021)
34. Yan, W.: Computational Methods for Deep Learning: Theoretic, Practice and Applications. Springer (2021)
35. Yan, W.: Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer (2019)
36. Gu, Q., Yang, J., Kong, L., Yan, W., & Klette, R.: Embedded and real-time vehicle detection system for challenging on-road scenes. Optical Engineering 56 (6), 06310210 (2017)
37. Ming, Y., Li, Y., Zhang, Z., & Yan, W.: A survey of path planning algorithms for autonomous vehicles. International Journal of Commercial Vehicles (2021).
38. Shen, D., Xin, C., Nguyen, M., & Yan, W.: Flame detection using deep learning. International Conference on Control, Automation and Robotics (2018)

39. Xin, C., Nguyen, M., & Yan, W.: Multiple flames recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 296-307 (2020)
40. Luo, Z., Nguyen, M., & Yan, W.: Kayak and sailboat detection based on the improved YOLO with Transformer. ACM ICCCV (2022)
41. Le, R., Nguyen, M., & Yan, W.: Training a convolutional neural network for transportation sign detection using synthetic dataset. IEEE IVCNZ (2021)
42. Alexey, B., ChienYao, W., Mark, L.: YOLOv4: Optimal speed and accuracy of object detection. Image and Video Processing (2020)
43. Chuyi, L. et, al. YOLOv6: A single-stage object detection framework for industrial applications. Computer Vision and Pattern Recognition (2022)
44. Chienyao, W., Alexey, B., Mark, L.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Computer Vision and Pattern Recognition (2022)
45. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition, 7132-7141 (2018).
46. Cao, Y. T., Wang, J. M., Sun, Y. K., & Duan, X. J. Circle marker based distance measurement using a single camera. Lecture Notes on Software Engineering, 1(4), 376 (2013)
47. Pan, C., Liu, J., Yan, W., Zhou, Y.: Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing (2021)
48. Pan, C., Yan, W.: Object detection based on saturation of visual perception. Multimedia Tools and Applications 79 (27-28), 19925-19944
49. Pan, C., Yan, W.: A learning-based positive feedback in salient object detection. IEEE IVCNZ (2018)
50. Shen, Y., Yan, W.: Blind spot monitoring using deep learning. IEEE IVCNZ (2018)
51. Zheng, K., Yan, W., Nand, P.: Video dynamics detection using deep neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence (2017)
52. An, N., Yan, W.: Multitarget tracking using Siamese neural networks. ACM Transactions on Multimedia Computing, Communications and Applications (2021)
53. Leslie, M., David, C., Sarah, L., Delphi, H., Teresa, M., Andrea, M., Lilliana, R., Francis, I., David, W., Sue, B.: Identification of the MuRF1 skeletal muscle ubiquitylome through quantitative proteomics (2021)
54. Xinyu, Z., Hongbo, G., Jianhui, Z. H. A. O., & Mo, Z. H. O. U.: Overview of deep learning intelligent driving methods. Journal of Tsinghua University (Science and Technology), 58(4), 438-444.(2018)
55. Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 37(3), 362-386. (2020)
56. Muhammad, K., Ullah, A., Lloret, J., Del Ser, J., & de Albuquerque, V. H. C.: Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 22(7), 4316-4336.(2020)
57. Mozaffari, S., Al-Jarrah, O. Y., Dianati, M., Jennings, P., & Mouzakitis, A. Deep learning-based vehicle behaviour prediction for autonomous driving applications: A review. IEEE Transactions on Intelligent Transportation Systems, 23(1), 33-47.(2020)
58. Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M. A., Cao, D., & Li, J.: Deep learning for lidar point clouds in autonomous driving: A review. IEEE Transactions on Neural Networks and Learning Systems, 32(8), 3412-3432.(2020)