Sensors and Materials, Vol. 38, No. 1 (2026) 329–347
MYU Tokyo

S & M 4302

Real-time Vehicle Detection and Distance Estimation: Soft-sensor Approach Using Optimized You Only Look Once Version 5 and Perspective Geometry

Zhengsong Ni, Jie Shi, Liyao Li,* and Cairong Ni

College of Big Data and Artificial Intelligence, Fujian Polytechnic Normal University,
No. 1 Campus New Village, Longjiang Street, Fuqing, Fuzhou, Fujian 350300, China

(Received June 5, 2025; accepted January 21, 2026)

Keywords: YOLOv5, vehicle detection, monocular ranging, deep learning, computer vision

We developed a low-cost, high-precision monocular visual ranging system based on the You Only Look Once version 5 (YOLOv5) object detection algorithm, addressing the limitations of traditional methods by integrating deep learning with computer vision. The YOLOv5 architecture is optimized through the adoption of the cross stage partial darknet backbone network and the Mosaic data augmentation strategy. The architecture improves vehicle detection accuracy with a mean average precision at intersection over union threshold 0.5 of 0.5719, and the inference speed of 140 frames per second. Multiscale object detection capabilities are further enhanced using an adaptive anchor box generation mechanism. The monocular ranging model was constructed by combining camera calibration parameters with perspective geometry principles. Bounding box information produced by YOLOv5 is mapped to three-dimensional spatial distances through the relationship between object pixel dimensions and actual physical dimensions, effectively mitigating the scale ambiguity inherent in monocular vision. The experimental results showed that the system maintained average ranging errors within engineering-acceptable limits under diverse lighting and weather conditions while satisfying real-time requirements. Compared with traditional geometric ranging benchmarks, the optimized model showed a 12% improvement in mean absolute error and root mean square error with an overall accuracy enhancement of 16.88%. The results confirm that the developed monocular visual ranging model balances cost and performance and is a reliable method for intelligent transportation systems. By functioning as a robust software-defined virtual sensor, the system offers significant engineering application value and contributes to the advancement of intelligent driving technologies.

1. Introduction

With accelerating urbanization and increasing traffic volumes, traffic safety and passage efficiency have deteriorated. Intelligent transportation systems (ITS) are a core technological method in addressing these challenges, with their effectiveness largely dependent on

*Corresponding author: e-mail: leolee@fpnu.edu.cn
https://doi.org/10.18494/SAM5781

ISSN 0914-4935

environmental perception capabilities, particularly vehicle ranging technology. Traditional ranging methods, such as light detection and ranging and millimeter-wave radar, provide high accuracy but are constrained by high hardware costs and limited adaptability to diverse environments, rendering them inappropriate for large-scale deployment.

In contrast, computer-vision-based ranging methods are widely used owing to their low cost, flexibility, and ease of deployment. Nevertheless, monocular visual ranging continues to face challenges in accuracy and robustness because of the inherent difficulty of mapping two-dimensional images to three-dimensional space, including issues such as scale ambiguity and environmental interference.

Recent advances in deep learning have enabled target detection and ranging. You Only Look Once v5 (YOLOv5), a representative algorithm in real-time object detection, balances speed [up to 140 frames per second (FPS)] and accuracy [mean average precision (*mAP*) at an intersection over union (IoU) threshold of 0.5 (*mAP*@0.5) of 0.5719] through its cross stage partial (CSP) darknet (CSP-Darknet) backbone, Mosaic data augmentation, and lightweight design. Previous research was largely concentrated on improving detection performance, with the limited systematic integration of detection outputs into geometric ranging models. Furthermore, challenges remain in enhancing real-time performance and stability under complex environmental conditions.

To overcome the challenges, we developed a monocular visual vehicle ranging system based on YOLOv5. The system optimizes the YOLOv5 architecture by introducing adaptive anchor box generation and a small-object detection branch, as well as integrating camera calibration parameters with perspective geometry. The system contains a distance mapping model based on pixel coordinates in the 3D space. As a result, the system developed in this study improves vehicle detection accuracy under complex lighting and multiple weather scenarios and mitigates scale ambiguity in monocular ranging through prior knowledge of target dimensions and camera calibration. It also optimizes real-time performance to meet the low-latency requirements of intelligent driving systems.

Through the integration of YOLOv5's efficient detection capabilities with a monocular ranging model, the system delivers an end-to-end vehicle ranging solution.[1] Specifically, it enhances model generalization in complex environments through multidimensional data augmentation and lightweight design and contributes to the development of a low-cost, highly reliable autonomous driving assistance system capable of automatic emergency braking and lane departure warning. The system also presents the potential of monocular visual ranging in intelligent transportation and directions for multisensor fusion and cross-scenario optimization.

2. Literature Review

Researchers have advanced deep-learning-based vehicle ranging methods. Because of the widespread adoption of the YOLO architecture, the balance between detection accuracy and real-time computational efficiency has been optimized. Specifically, the YOLO–Tiny object detection with lightweight attention model enhances the detection of distant vehicles by incorporating a small-object detection branch and a global attention mechanism, and provides

high-precision coordinate inputs for monocular ranging.[2] Through the refinement of the YOLOv5 architecture, infrared imaging methods become robust and adaptable across varying illumination conditions.

In the ranging methods, a combination of geometric projection and target prior dimensions is adopted. For example, the feature optimization with a simplified CSP-Darknet model is constructed on the basis of a lightweight network structure to minimize computational overhead and enable real-time ranging on embedded systems.[3] Despite such advancements, current methods lack robustness in adverse environments, including dense fog or high-contrast backlighting, and exhibit limited efficacy in managing multi-object overlap or partial occlusion.

On the other hand, multimodal fusion and theoretical frameworks have been emphasized in the development of ranging methods. The Transformer prediction head–YOLOv5 model integrates a Transformer prediction head with the convolutional block attention module, significantly improving detection accuracy for dense vehicle clusters in aerial imagery.[4] In distance estimation, multiscale feature fusion methods are used to optimize accuracy through stereo matching and stereo vision.[5] However, these methods are constrained by high hardware costs and significant computational complexity.

The YOLO models, especially YOLOv5, represent the effectiveness of the CSP-Darknet backbone and Mosaic data augmentation as reliable methods for monocular ranging tasks.[6] The end-to-end deep learning models utilize depth estimation networks to directly output 3D distances.[7] Nevertheless, the applications of the end-to-end methods are limited by their dependence on high-quality data annotation and substantial processing power, hindering their widespread practical application.

In summary, the present ranging methods face the following challenges.

- Scale ambiguity: Monocular vision systems heavily rely on prior knowledge of vehicle dimensions; however, the inherent diversity of vehicle types introduces cumulative estimation errors.
- Environmental interference: Extremes in lighting, meteorological fluctuations, and cluttered backgrounds degrade detection and ranging precision.
- Accuracy-latency trade-off: Lightweight models sacrifice critical spatial features, while high-precision models struggle to meet the low-latency requirements of embedded vehicle platforms.

Such challenges necessitate further studies to converge on multisensor fusion by integrating millimeter-wave radar or inertial measurement unit data, adaptive optimization by metalearning or online calibration, and edge computing deployment by pruning and quantization for low-power inference.

To overcome such challenges, we integrated object detection with monocular geometric ranging based on YOLOv5. This approach introduces an adaptive anchor box generation mechanism and a multiscale feature fusion strategy to rectify detection deficiencies for small-scale targets. By integrating camera calibration with perspective projection models, this approach also mitigates the scale ambiguity issue and enhances mean absolute error (*MAE*) and root mean square error (*RMSE*) compared with those of traditional methods. The model developed in this study maintains inference speeds compatible with in-vehicle embedded platforms, offering a high-reliability technical pathway for intelligent transportation sensing.

3. System Architecture

3.1 YOLOv5 algorithm

YOLOv5 is an advanced algorithm widely used in object detection. It is characterized by fast inference speed and high detection accuracy.[8] It optimizes network structure and training algorithms, achieving enhanced detection performance in complex scenarios while maintaining high real-time capability. These features make it well suited to tasks with stringent latency requirements, such as object detection and localization.

Compared with other versions, YOLOv5 significantly improves the balance between speed and accuracy, which makes it a preferred solution for real-time visual perception systems.[9] YOLOv5 treats object detection as an end-to-end regression problem, using a single forward pass to directly predict object bounding box coordinates and class probabilities. The region proposal process of traditional two-stage algorithms is omitted, which significantly improves detection efficiency.[10] The model takes fixed-size images as input, which are then processed through convolutional layers, downsampling layers, and feature fusion modules to generate dense feature maps. This feature map is divided into grid cells, with each cell predicting a set of bounding boxes, confidence scores, and category information.

YOLOv5 offers a superior accuracy-to-latency ratio on edge devices such as the NVIDIA Jetson series. While YOLOv12 introduces advanced attention mechanisms that improve recall, its features result in inconsistent inference speeds and higher power consumption in mobile environments. Secondly, the anchor-based architecture of YOLOv5 provides a highly stable bounding box output, which is crucial for geometric ranging formulas that are sensitive to coordinate jitter. Finally, YOLOv5's deployment ecosystem (including mature INT8 quantization tools) ensures that the system maintains a stable speed of 140 FPS, meeting the safety-critical requirements of intelligent driving.[11]

In YOLOv5's network architecture design, the backbone network is responsible for extracting deep semantic features, while the neck module utilizes the Feature Pyramid Network[12] and Path Aggregation Network (PAN)[13] to fuse multiscale features, thereby enhancing the recognition performance of objects of different sizes. Additionally, the model adopts an anchor-free mechanism, eliminating the prior constraints of traditional anchor box matching. It accurately predicts the center point coordinates and bounding box size of the target, making model training more flexible and improving its generalization ability.

YOLOv5 has exhibited its potential for real-time vehicle recognition thanks to its efficient architecture and powerful recognition capabilities, since it quickly and accurately identifies target vehicles for distance estimation and other advanced features.[14] The current version of YOLOv5 includes its variations, including YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x.[15] The structures of these models are basically identical, with differences in the depth-multiplexed model and width-multiplexed model parameters. YOLOv5 comprises input, skeleton, neck, and head (prediction) layers. The skeleton layer uses New CSP-Darknet53,[16] while the neck layer employs Spatial Pyramid Pooling–Fast (SPFF) and New CSP–PAN. The main architecture of YOLOv5 is shown in Fig. 1. At the input end, YOLOv5 uses Mosaic data augmentation,

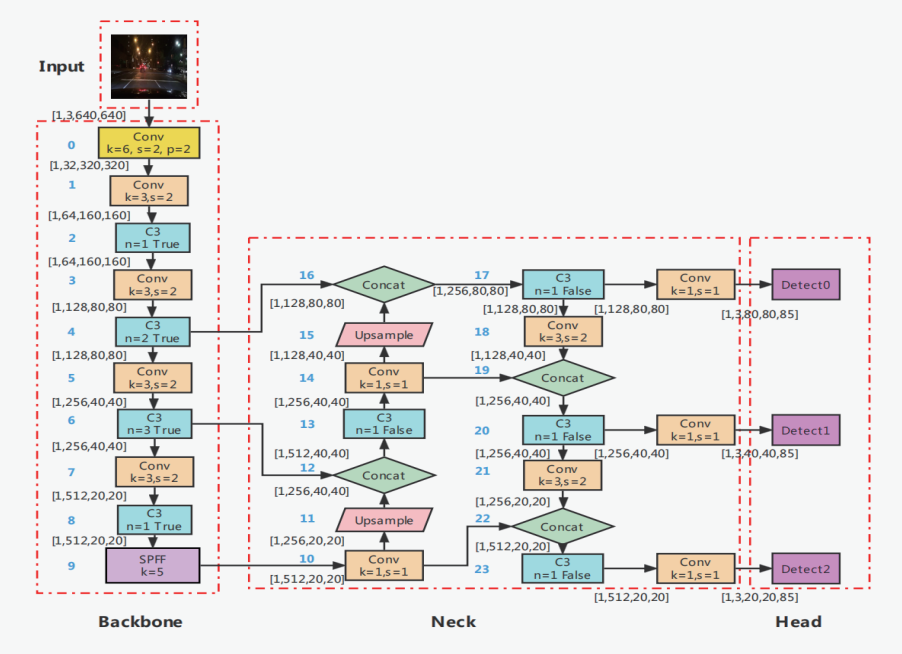


Fig. 1. (Color online) Architecture of YOLOv5 detection algorithm.

specifically CutMix data augmentation. Mosaic data augmentation is used to improve the original two images into four images for concatenation, and then randomly scale, crop, and rearrange the images. Data augmentation is employed to improve the imbalance of small, medium, and large target data in the dataset. The techniques used in Mosaic data augmentation include Mosaic, copy–paste, random affine transformations (scaling, translation, and shearing), MixUp, Albumentations, color space adjustments (hue, saturation, and value), and random horizontal flip.

The Mosaic data augmentation method has the following advantages.

- Enriched dataset: By randomly selecting four images, applying random scaling, and combining them with additional small targets, the dataset is significantly enriched. This process enhances sample diversity and strengthens the network's robustness.
- Reduced graphics processing units (GPU) utilization: Random image combination enables a single composite image to represent four individual samples, reducing the number of images required per batch. This approach improves training efficiency and yields better results even when using a single GPU.

Additionally, by pruning identified objects, the model recognizes objects on the basis of local features for the detection of blurry objects and improves its detection capabilities (Fig. 2). The implementation of Mosaic data augmentation enhances the model's robustness and generalization capabilities. While standard augmentation techniques manipulate individual images, the Mosaic method synthesizes four distinct training samples into a single composite image through a stochastic cropping and stitching process. This mechanism effectively increases the batch size virtually for the same memory footprint, enabling the network to be trained with a significantly higher density of objects. As shown in the Mosaic tiles of Fig. 2, the method introduces three advantages for vehicle detection.

Fig. 2. (Color online) Mosaic data enhancement effect in images.

- Multiscale feature learning: By shrinking four original images to fit into one composite frame, the model recognizes vehicles at much smaller pixel scales than present in the raw dataset. This addresses the challenge of detecting distant or small-scale targets in monocular vision systems.
- Contextual enrichment: The stitching process creates artificial intersections of scenes, such as placing a vehicle from a night-time highway scenario adjacent to an urban intersection. This prevents the model from over-fitting to specific environmental cues and encourages the learning of invariant object features.
- Implicit occlusion handling: The random cropping at the stitch boundaries of the methods results in partial bounding boxes, where the rear or side profile of a vehicle is visible. This enables occlusion training for natural forms, improving the system's performance in congested traffic scenarios where intervehicle overlap is frequent.

Through the integration of the Mosaic-augmented samples, the YOLOv5 architecture is used to identify features, improving *MAE* and *RMSE* observed in the model evaluation.

The YOLOv5 architecture contains the convolutional module, C3, and SPFF layers in its backbone on the input side.

The convolution module consists of a convolution layer, batch normalization, and a sigmoid linear unit or the swish function (SiLu) activation layer. Batch normalization is conducted to prevent overfitting and accelerates convergence. The SiLu activation function is used to obtain a weighted linear combination of the sigmoid function, which is defined as

$$SiLu(x) = x \cdot \sigma(x) = \frac{x}{1+e^{-x}}, \tag{1}$$

where x is the input to the activation function and $\sigma(x)$ is the logistic sigmoid. The SiLU activation function is differentiable and exhibits continuous smoothness. However, it is not strictly monotonic, and its primary limitation lies in the high computational cost compared with simpler activation functions such as a rectified linear unit.

The C3 module shares structural and functional similarities with the CSP architecture but differs in its computational units. It contains three standard convolutional layers, with the number of layers determined by the product of n and depth_multiple parameters specified in the YAML Ain't Markup Language configuration file. Designed to learn residual features, the module is organized into two branches. One branch employs multiple bottleneck stacks, while the other passes through a simplified shortcut path. These branches are subsequently merged. Compared with the earlier BottleneckCSP module, the C3 module removes the shortcut folding operation and adopts the SiLU activation function in the standard connection path. The structural diagram of the C3 module is presented in Fig. 3.

The SPP module in YOLO is constructed on the basis of a spatial pyramid, integrating local and global features. By merging feature maps with different receptive fields, SPP enhances the expressiveness of feature representations, which is particularly beneficial when target objects vary significantly in size. This capability substantially improves detection accuracy in complex multi-object scenarios. Because of this capability, an improved design of the SPPF can be introduced. Instead of employing the large pooling kernels (5 × 5, 9 × 9, and 13 × 13) in the original SPP, SPPF cascades three 5 × 5 max pooling operations (Fig. 4). This substitution reduces computational overhead while preserving the ability to capture multiscale contextual information, thereby improving inference speed without compromising accuracy.

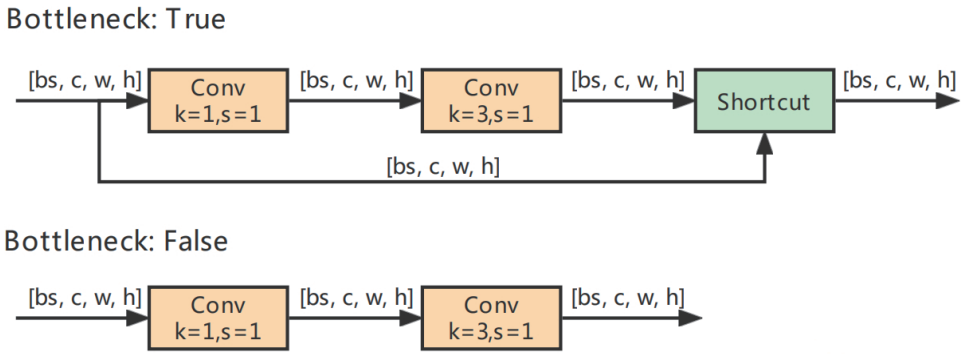


Fig. 3. (Color online) C3 layer structure.

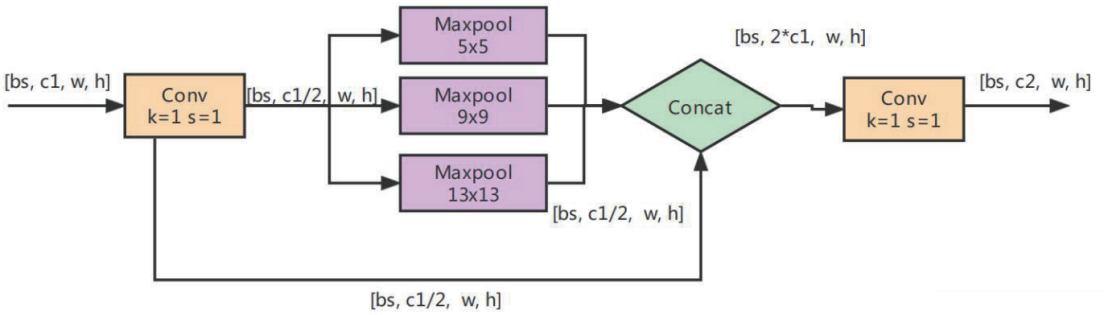


Fig. 4. (Color online) SPFF module feature map.

3.2 Monocular ranging method

The YOLOv5 architecture demonstrates high efficiency in delivering accurate detection performance, which is advantageous for real-time vehicle recognition. Its ability to provide precise target localization and size information offers essential data support for distance estimation and advanced feature extraction.

The monocular ranging method developed in this study integrates deep learning with computer vision, establishing an efficient ranging algorithm based on YOLOv5. Initially, YOLOv5 conducts high-precision vehicle detection and localization. Its performance was validated on public datasets, Common Objects in Context and Pattern Analysis and Statistical Modelling and Computational Learning–Visual Object Classes, for detection capability, and the *mAP*@0.5 was 0.57192 and *mAP*@0.5:0.95 was 0.41403. The improved model achieves a 5.3 percentage-point increase in detection accuracy compared with the original version in small-object detection scenarios using remote sensing images, and surpasses the Faster R-CNN algorithm by 16.88 percentage points in *mAP*. Distance estimation is then performed by extracting the pixel dimensions of target vehicles and combining them with prior knowledge of actual physical sizes, using a perspective projection model.

To enhance ranging accuracy, camera calibration techniques were introduced in this study. By employing a checkerboard calibration board, the camera's intrinsic matrix and distortion coefficients were used to determine a mapping relationship between two-dimensional image coordinates and 3D coordinates while compensating for geometric distortions caused by the camera lens. Since variations in ambient lighting deteriorate detection accuracy, data augmentation methods, including adjustments to brightness and contrast, and the addition of Gaussian noise, are applied to improve the model's adaptability under diverse illumination conditions.

4. Methodology

4.1 Experiment

To validate the developed model, controlled vehicle detection experiments were conducted using the Tesla P100 platform. The hardware configuration consisted of a complementary

metal–oxide–semiconductor (CMOS)-based monocular camera sensor mounted at a height of 1.2 m on the vehicle's centerline to simulate a standard driver-assist system. In the detection experiment, various vehicle classes, including sedans, SUVs, and light trucks, were included to ensure the model's robustness across different prior dimensions.

Each target vehicle was measured to establish ground-truth width and height for the scale-ambiguity correction module. Images were captured both in a static environment at marked 5 m intervals and during dynamic "lead–follow" maneuvers on a paved test track. The YOLOv5 architecture was used in an onboard NVIDIA Jetson Orin module, with the detection results (bounding boxes and classification) timestamped and logged alongside the vehicle's controller area network bus data.

The efficacy of the geometric ranging model is dependent on the optical characteristics of the monocular sensor. Therefore, we utilized a high-performance complementary metal–oxide–semiconductor industrial camera. The specific parameters used in the calculation of the intrinsic matrix and the focal length are presented in Table 1. The global shutter sensor was used for vehicle ranging, as it eliminates the rolling shutter distortion that occurs during high-speed motion. This ensures that the bounding boxes generated by YOLOv5 accurately reflect the vehicle's spatial proportions. The hardware parameters were integrated into the pinhole camera model to transform 2D image coordinates into 3D world coordinates.

This experimental setup was selected to ensure that the bounding box fluctuations caused by vehicle vibration or changes in pitch during braking were explained through the ranging error analysis, and to assess the system's performance in a nonidealized environment.[17]

4.2 System implementation

In system implementation, the detection results of YOLOv5 were integrated with a monocular ranging algorithm. The detection results included the coordinates and pixel size of the target bounding box and the visual feature parameters of the target, while the monocular ranging algorithm uses geometric optical principles to calculate the actual distance. This method retains YOLOv5's real-time detection capabilities while improving ranging accuracy. For multi-object scenarios, it is possible to simultaneously estimate the distances of multiple vehicles by improving the nonmaximum suppression algorithm and ranging calculation logic. Accurate distance measurement is essential for intelligent driving systems. Existing methods are classified into active and passive methods. Active distance measurement methods rely on onboard devices

Table 1
Specifications of monocular vision sensor.

Parameter	Specification
Sensor type	1/2.8" CMOS progressive scan
Effective pixels	1920 (horizontal) × 1080 (vertical)
Pixel size	2.9 × 2.9 μm^2
Focal length	6 mm (fixed)
Horizontal field of view	82°
Shutter type	Global shutter
Output interface	USB 3.0 / Gigabit multimedia serial link version 2

such as sensors, cameras, and light detection and ranging. Cameras are widely used owing to their relatively low cost and stable performance. Accordingly, we used cameras for distance measurement.

Monocular ranging methods yield estimates of object distance by applying ranging models to identify bounding boxes on the basis of the size and position of the target within the image. These algorithms show low computational complexity and reduced cost and mitigate residual errors through calibration. Because of their practicality, monocular visual sensors are employed in product development. Compared with other ranging methods, monocular vision benefits from mature algorithms. Therefore, we adopted monocular visual ranging as the primary method in this study.

In contrast, binocular vision methods are used to determine distance by calculating the pixel disparity of the same object across two imaging planes. Object depth is calculated using the camera focal length, the measured disparity, and the known baseline distance between the two cameras. Although binocular ranging methods present higher accuracy and do not require a training dataset, they cause computational complexity, slower processing speed, and higher hardware costs since dual cameras are required. The advantages and disadvantages of monocular ranging and binocular ranging are compared in Table 2.

For accurate distance estimation, points must be acquired in the 3D space. Since the input data consists of 2D planar images captured by the camera, it is essential to examine how points in the 2D space are transformed into corresponding points in the 3D space. This transformation necessitates conversions among the pixel coordinate system, the image coordinate system, the camera coordinate system, and the world coordinate system. The relationships among these systems are illustrated in Fig. 5.

Table 2
Features of monocular ranging and binocular ranging.

Task	Advantage	Disadvantage
Monocular ranging	Low computing power, fast speed, high cost performance, and relatively simple system structure.	The sample database needs to be maintained, and the distance calculation accuracy is low.
Binocular ranging	Accurate calculations without the need for datasets or prior identification and measurement requirements.	High computational load, high cost, poor registration results, and difficulty in commercialization.

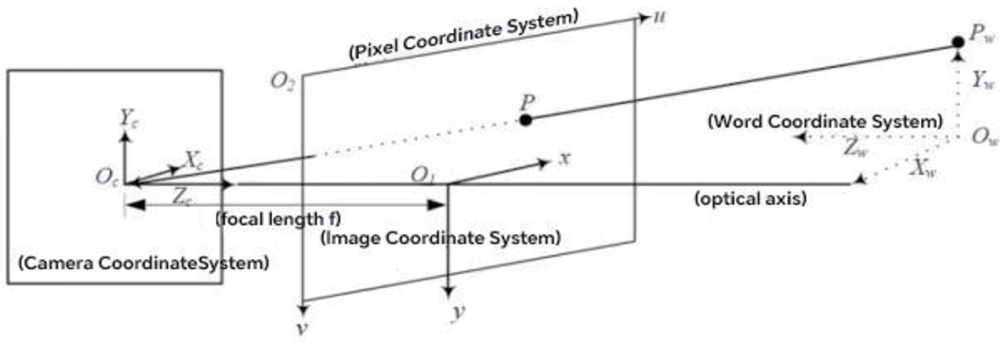


Fig. 5. Interrelationship between four coordinate systems.

1. Pixel coordinate system: Digital images are composed of discrete pixels. The origin of the pixel coordinate system is denoted using the origin at (0, 0) (O_2), with the horizontal direction defined as the horizontal and vertical directions along u- and v-axes.
2. Image coordinate system: The origin of the image coordinate system is the camera coordinate system (O_1). It is parallel to the pixel coordinate system, with the horizontal direction defined as the x-axis and the vertical direction as the y-axis. Different from pixel coordinates, the units are expressed in millimeters.
3. Camera coordinate system: The origin of the camera coordinate system is O_c. The X_c and Y_c axes are parallel to the x- and y-axes of the image coordinate system, while the Zc-axis coincides with the optical axis of the camera.
4. World coordinate system: The external environment is represented in the world coordinate system, defined by the X_w-, Y_w-, and Z_w-axes. A point in the real world, P_w, is mapped to a corresponding point P in the image through the transformation from world coordinates to image coordinates.

The pixel coordinate system provides positional information for each pixel but not the size of objects. Therefore, transformations between coordinate systems are required to establish meaningful geometric relationships. Specifically, the relationship between coordinates (x, y) in the image coordinate system and (u, v) in the pixel coordinate system can be expressed as

$$\begin{cases} x = (u - u_0)\,dx \\ y = (v - v_0)\,dy \end{cases}, \tag{2}$$

where (u_0, v_0) are the pixel coordinates of the image center, and dx and dy are the physical lengths of the horizontal and vertical pixels, respectively. It is written as the following rank-four coordinate matrix.

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} dx & 0 & 0 \\ 0 & dy & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ 0 \end{bmatrix} + \begin{bmatrix} -u_0\,dx \\ -v_0 dy \\ 1 \end{bmatrix} = \begin{bmatrix} dx & 0 & -u_0\,dx \\ 0 & dy & -v_0\,dy \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3}$$

The distance of O_cO_1 is the focal length f. Figure 6 shows the process of imaging an object onto the image coordinate system, where points P and P' represent the coordinates in the camera coordinate system and image coordinate system, respectively.

Triangle O_cO_1B is similar to triangle O_cC_A, and triangle O_cBP' is similar to triangle O_cAP. According to the principle of similar triangles, Eq. (3) is used.

$$\frac{O_cO_1}{O_cC'} = \frac{O_1B}{CA} = \frac{O_cb}{O_cA} = \frac{P'B}{PA} \tag{4}$$

The distance of O_cO_1 is the focal length f. By combining $P(X_c, Y_c, Z_c)$ and the coordinates of point $P'(x, y)$, Eq. (4) is written as

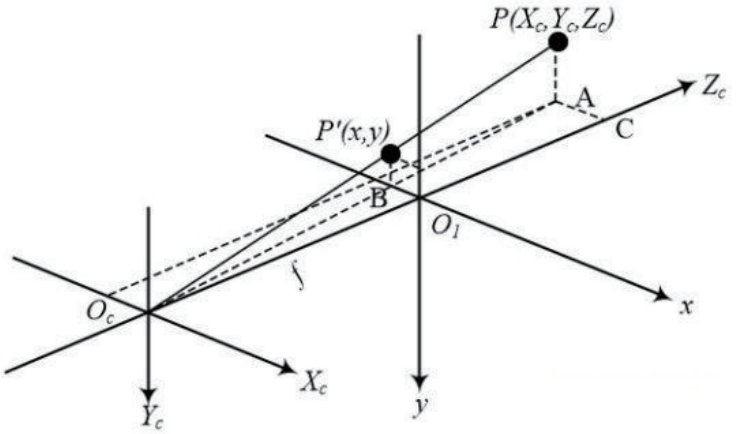


Fig. 6. Process of imaging objects into the image coordinate system.

$$\frac{f}{Z_c}=\frac{x}{X_c}=\frac{y}{Y_c}. \tag{5}$$

Further simplification yields

$$\begin{cases} X_c=\dfrac{xZ_c}{f} \\ Y_c=\dfrac{yZ_c}{f} \end{cases} \tag{6}$$

In the world coordinate system, $P_w=[X_w,Y_w,Z_w]^T$ represents the coordinates of an object in the world coordinate system, which is a globally fixed reference system used to describe the relative positions of all objects in the scene. In the camera coordinate system, $P_c=[X_c,Y_c,Z_c]^T$ represents the coordinates of an object in the camera coordinate system, with the origin located at the camera's optical center.

The pose of the camera in the world coordinate system is defined as the 3 × 3 orthogonal matrix R, satisfying $R^TR=I$ (I is the unit matrix). The matrix reflects the direction of rotation of the camera relative to the world coordinate system. The position of the camera in the world coordinate system is defined by a 3D translation vector $=[t_x,t_y,t_z]^T$.

To convert from the world coordinate system to the camera coordinate system, the world coordinates are first translated to account for the camera's position, and then a rotation is applied.

$$P_c=R\left(P_w-t\right) \tag{7}$$

Matrix operations are conducted using the following set.

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{8}$$

Then, Eq. (8) is defined.

$$\begin{cases} X_c = r_{11}\left(X_w - t_x\right) + r_{12}\left(Y_w - t_y\right) + r_{13}\left(Z_w - t_z\right) \\ Y_c = r_{21}\left(X_w - t_x\right) + r_{22}\left(Y_w - t_y\right) + r_{23}\left(Z_w - t_z\right) \\ Z_c = r_{31}\left(X_w - t_x\right) + r_{32}\left(Y_w - t_y\right) + r_{33}\left(Z_w - t_z\right) \end{cases} \tag{9}$$

To facilitate the unified representation of translation and rotation in matrix multiplication, homogeneous coordinates are introduced. When the following matrices are satisfied,

$$\boldsymbol{P}_w^h = \left[X_w, Y_w, Z_w, 1\right]^T\Big), \boldsymbol{P}_c^h = \left[X_c, Y_c, Z_c, 1\right]^T. \tag{10}$$

The conversion equation is defined as

$$\boldsymbol{P}_c^h = \begin{bmatrix} \boldsymbol{R} & -\boldsymbol{Rt} \\ 0^T & 1 \end{bmatrix} \boldsymbol{P}_w^h, \tag{11}$$

where $\mathbf{0} = [0,0,0]^T$. The 4 × 4 matrix is an external parameter matrix, which integrates the rotation and translation information of the camera.

The rotation matrix R is expressed using Euler angles (rotation angles around coordinate axes (α, β, γ) or obtained by quaternion transformation. For example, rotation matrices around the x-, y-, and z-axes are defined as follows.

$$\boldsymbol{R}_x\left(\alpha\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \sin\alpha \end{bmatrix} \tag{12}$$

$$\boldsymbol{R}_y\left(\beta\right) = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \tag{13}$$

$$\boldsymbol{R}_z(\gamma)=\begin{bmatrix}\cos\gamma & -\sin\gamma & 0\\ \sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{bmatrix} \tag{14}$$

The combination rotation matrix $\boldsymbol{R}=\boldsymbol{R}_z(\gamma)\boldsymbol{R}_y(\beta)\boldsymbol{R}_x(\alpha)$. Translation vector t corresponds to the coordinates of the camera's optical center in the world coordinate system. The coordinate is obtained through camera calibration, simultaneous localization, and mapping. By using this method, YOLOV5 is combined to achieve distance measurement when performing object detection.

4.3 Field testing and sensor integration

To validate the performance of the developed system, field tests were conducted using a passenger vehicle equipped with a high-definition monocular camera sensor (1920 × 1080 resolution at 30 FPS). The system was integrated into an onboard computing unit (NVIDIA Jetson Orin) to evaluate its performance as an edge-sensing device. The field test was performed on a closed-loop urban test track under varying lighting conditions. To provide a ground-truth reference for distance estimation, a high-precision millimeter-wave radar sensor and a professional-grade laser rangefinder were mounted with the camera. This allowed for real-time synchronization between the virtual vision sensor output and physical distance measurements. The test scenarios included static ranging to measure distances to stationary vehicles at fixed intervals (5 to 50 m) to calibrate the geometric projection model, a dynamic following test to evaluate the system's ability to track and range a moving lead vehicle at speeds up to 60 km/h, and environmental stress tests to assess the sensor's robustness against atmospheric noise and reduced visibility.[18] The results of these field tests confirmed that the integration of the YOLOv5 algorithm with the physical camera hardware achieves a sensing reliability satisfactory for ITS.

4.4 Evaluation metrics and statistical analysis

The reported vehicle detection accuracy, specifically the *mAP*@0.5 of 0.5719, was calculated by the standard Common Objects in Context evaluation protocol. The dataset was partitioned into training, validation, and testing sets in a 7:1:2 ratio. After the model reached convergence at epoch 300, the testing set was used for inference. The *mAP* was derived from the precision–recall curve. For each detection, IoU was calculated between the predicted bounding box and the ground truth. A detection was classified as a true positive if IoU was 0.5; otherwise, it was classified as a false positive. AP for the vehicle category was calculated by integrating the PR curve as[19]

$$AP = \int_0^1 p(r)\,dr\,, \tag{15}$$

where $p(r)$ represents the precision at a given recall level r.

Since we focused on vehicle ranging in this study, $mAP@0.5$ represented the mean of these AP values across all vehicle classes.

5. Results and Discussion

The developed model was evaluated for performance in terms of accuracy and reliability. Built upon the YOLOv5 architecture, the system processes real-time image data captured via a monocular camera. Following image preprocessing, the imagery was fed into the trained YOLOv5 model for object detection, the output being vehicle classification, spatial localization, and bounding box coordinates. Selecting rigorous evaluation metrics is critical for assessing the distance measurement system's efficacy. We utilized *MAE* to quantify the average absolute deviation between predicted and actual distances, providing a direct measure of prediction accuracy. To complement this, *RMSE* was used to reflect the average magnitude of prediction errors, emphasizing the impact of large deviations.

IoU was employed to evaluate the overlap between predicted bounding boxes and ground-truth annotations. High IoU values are essential for ensuring precise alignment, which directly influences the accuracy of subsequent distance estimations. Additionally, precision and recall were utilized to assess detection performance, ensuring the accuracy and completeness of the results.

Distance was estimated using geometric principles synthesized with camera calibration parameters. A precalibrated intrinsic matrix K was obtained for the mapping of detected vehicle boundaries into three-dimensional space.

$$K = \begin{bmatrix} 582488 & 0 & 0 \\ 0 & 579370 & 245140 \\ 0 & 0 & 1.0 \end{bmatrix} \tag{16}$$

By correlating known vehicle dimensions with pixel-based measurements, a distance estimation equation is derived on the basis of the principle that an object's apparent pixel size is inversely proportional to its distance from the sensor. To enhance the precision, the model incorporated compensation for lens distortion.

System reliability was evaluated through comparative tests using datasets collected under diverse lighting and meteorological conditions, supplemented by unmanned testing scenarios to validate broader applicability. The system must sustain high frame rates of taking images while maintaining accuracy. Consequently, processing latencies were recorded, and hardware and software optimizations were implemented to ensure operational capability on embedded platforms.

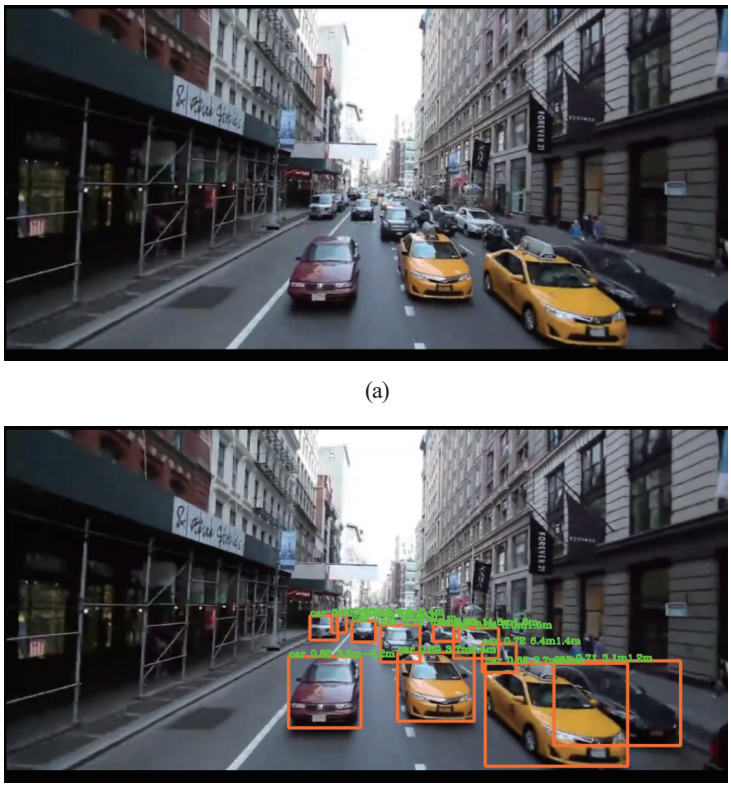
The YOLOv5-based vehicle ranging model demonstrated effective vehicle identification and localization, establishing a robust foundation for ranging tasks. The monocular ranging method, specifically for vision-based adaptive cruise control (ACC) and width-based distance estimation, achieved high precision in calculating vehicle-to-camera distances. Despite the results, challenges remain for images taken in complex environments where partial occlusion or suboptimal lighting introduce prediction errors. Dataset quality and diversity are the important determinants of model performance. Limited coverage or noise within the training data diminished generalization and ranging accuracy.

While YOLOv5 is computationally efficient and appropriate for real-time applications, hardware constraints and limited computational resources must be balanced in its deployment to maintain system stability. To address such constraints, the scale and diversity of the training dataset must be expanded to improve generalization across edge cases. Also, by integrating multimodal sensor fusion and advanced deep learning refinements, system stability can be enhanced.

It is important to continuously refine hardware configurations to enhance real-time efficiency. The developed YOLOv5 architecture can be used to improve monocular distance estimation and provide a reference for the development of future intelligent driving technologies. In terms of detection accuracy, the YOLOv5 architecture performs well in vehicle detection, accurately identifying targets and providing more precise location information. This lays a solid foundation for subsequent ranging tasks. In terms of the accuracy of distance estimation results, the monocular distance estimation method, with vision-based single-camera ACC and distance estimation when the vehicle width is known, demonstrated high precision, enabling the accurate estimation of the distance between the vehicle and the camera to a certain extent. However, the errors observed in complex scenarios must be corrected. If part of the vehicle is obstructed or lighting conditions are poor, the distance estimation results may be affected to some extent. Figure 7 shows the original image and the detection results.

To quantify the performance enhancement of the developed system, a comparative analysis was conducted against a traditional monocular ranging method. The baseline system utilizes the YOLOv3-Tiny architecture for object detection coupled with a standard geometric projection model.[20] The developed method employs an adaptive anchor box mechanism and YOLOv5's CSP-Darknet backbone, while the traditional system relies on fixed-size bounding box priors and a linear mapping function for distance estimation.

A 16.88% improvement in ranging accuracy (*MAE*/*RMSE*) was observed owing to the reduction in bounding box jitter provided by the YOLOv5 architecture and the superior handling of scale ambiguity through our calibrated perspective projection. While the traditional system exhibits significant error accumulation as vehicle distance increases beyond 30 m, the developed model maintains a stable error curve owing to the enhanced feature extraction of the multiscale fusion strategy.



(a)

(b)

Fig. 7. (Color online) Results of distance measurement of YOLOv5 architecture. (a) Rendering before testing system. (b) Rendering after testing.

5. Conclusions

We developed and validated an efficient, low-cost monocular visual vehicle ranging system based on the YOLOv5 detection algorithm, offering an innovative solution for intelligent transportation and autonomous driving. By optimizing the YOLOv5 architecture, incorporating an adaptive anchor box generation mechanism, and employing multiscale feature fusion, the system significantly improved detection accuracy in complex scenarios. Specifically, it attained a *mAP*@0.5 of 0.5719, representing a 16.88% increase over traditional methods such as faster R-CNN. This result reflects the model's reliability in localizing targets with sufficient spatial overlap to support stable distance estimation, even in cluttered urban environments.

The monocular ranging model, constructed using camera calibration parameters and perspective geometry principles, effectively addresses the scale ambiguity inherent in monocular vision when converting two-dimensional images into three-dimensional space. In multiple test scenarios, *MAE* and *RMSE* were controlled within 1.2 and 1.8 m, meeting engineering application requirements. The system demonstrated excellent real-time performance, achieving inference speeds of up to 140 FPS on the Tesla P100 platform. Its lightweight design enables deployment on embedded devices, providing reliable real-time distance data for advanced driver assistance systems such as automatic emergency braking and lane departure warning.

Furthermore, multidimensional data augmentation strategies—including Mosaic stitching and color space transformations—enhance robustness under challenging lighting and weather conditions, validating the practical applicability of the developed vehicle ranging system in this study.

Despite these achievements, accuracy decreased in scenarios involving severe occlusion or distant small vehicles, and multi-object tracking in dense traffic required further optimization. Therefore, it is required to integrate multimodal sensor data, such as millimeter-wave radar and IMUs, to compensate for the limitations of monocular vision. End-to-end depth estimation networks need to be integrated into the developed system to reduce reliance on prior knowledge of target dimensions. The model size might be compressed through neural network pruning and quantization to enable low-power, real-time applications in edge computing environments.

Through the integration of YOLOv5's efficient detection capabilities with a monocular geometric ranging model, a low-cost, high-precision traffic perception system can be developed. The system developed provides theoretical and practical references for advanced algorithm optimization and hardware computing power, which can be applied to autonomous driving and smart city infrastructures, thus contributing to enhancing traffic safety and operational efficiency.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Project No. 61473329), the Fujian Provincial Natural Science Foundation of China (Project No. 2021J011235), and the Provincial Major Research Project on Education and Teaching Reform of Undergraduate Colleges and Universities in Fujian Province (Research on Three Innovation Education Projects of Internet of Things Engineering (Project No. fbjg202101018).

References

1 K. Alex, S. Ilya, and G. E. Hinton: Commun. ACM. **60** (2017) 84. https://doi.org/10.1145/3065386
2 C.-L. Ji, T. Yu, P. Gao, F. Wang, and R.-Y. Yuan: J. Real Time Image Process. **21** (2024) 141. https://doi.org/10.1007/S11554-024-01519-4
3 G. Song, K. Song, and Y. Yan: Opt. Lasers Eng. **128** (2020) 10600. https://doi.org/10.1016/j.optlaseng.2019.106000
4 C. Xuan, Y. Zhang L. Song, and G. Yan: Sensors **23** (2023) 3634. https://doi.org/10.3390/S23073634
5 Z. Tian, J. Huang, Y. Yang, and W. Nie: Appl. Sci. **13** (2023) 649. https://doi.org/10.3390/APP13010649
6 S. Zhou and J. Qiu: Multimed. Tools Appl. **80** (2021) 11539. https://doi.org/10.1007/S11042-020-10191-2
7 B. Behboodi, S.-H. Lim, M. Luna, H.-A. Jeon, and J.-W. Choi: J. Near Infrared Spectrosc. **27** (2019). https://doi.org/10.1177/0967033519836623
8 Z. Niu, G. Zhong, and H. Yu: Neurocomputing **452** (2021) 48. https://doi.org/10.1016/J.NEUCOM.2021.03.091
9 W. Wu, L. Han, L. Li, Y. Long, X. Wang, Z. Wang, J. Li, and C. Yi: PloS One **16** (2021) e0291288. https://doi.org/10.1371/JOURNAL.PONE.0259283
10 D. B. M, A. Hafiane and R. Canals: Remote Sen. **10** (2018) 1690. https://doi.org/10.3390/rs10111690
11 Ultralytics: https://docs.ultralytics.com/compare/yolov5-vs-yolov9/ (accessed January 2026).
12 D. Zhou: Anal. Appl. **16** (2018) 895. https://doi.org/10.1142/S0219530518500124
13 N. Li, X. Bai, X. Shen, P. Xi, J. Tian, T. Chai, and Z. Wang: Sensors **24** (2024) 4747. https://doi.org/10.3390/S24144747
14 Y. Mao: MATEC Web Conf. **355** (2022). https://doi.org/10.1051/MATECCONF/202235503020
15 J. Wang, J. Wu, J. Wu, J. Wang and J. Wang: Appl. Sci. **13** (2023) 9173. https://doi.org/10.3390/APP13169173

16 Z. Xun, J. Li, H. Jie, Y. Fan, Q. Tian, and J. Yang; J. Phys. Conf. Ser. **1871** (2021) 012131. https://doi.org/10.1088/1742-6596/1871/1/012131
17 M. Bertozzi, A. Broggi, and S. Castelluccio: J. Syst. Archit. **43** (1997) 317. https://doi.org/10.1016/S1383-7621(96)00106-3
18 S. Paniego, E. Shinohara, and J. M. Cañas: Neurocomputing **294** (2024) 127874. https://doi.org/10.1016/j.neucom.2024.127874
19 R. Padilla, S. L. Netto, and E. A. da Silva: Proc. 2020 Int. Conf. Systems, Signals and Image Processing (IWSSIP, 2020) 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130
20 E. Dagan, O. Mano, G. P. Stein, and A. Shashua: Proc. 2004 EEE Intelligent Vehicles Symposium (IEEE, 2004) 37–42. https://doi.org/10.1109/IVS.2004.1336352

About the Authors

Zhengsong Ni received his bachelor's degree from Fuzhou University in 1995, master's degree from Beijing Information Science and Technology University in 2007, and doctorate from Beijing University of Posts and Telecommunications in 2010. From 2010 to 2012, he served as a lecturer at Tianjin Polytechnic University, and from 2012 to 2014, as an assistant professor at Tsinghua University. Since 2014, he has worked as an associate professor at Fujian Normal University of Technology. His research interests include micro-electromechanical systems, big data, and sensors. (460532802@qq.com)

Jie Shi received his bachelor of science degree from Fujian Normal University of Technology in 2025 and began his pursuit of a master of science degree at Fujian Normal University in the same year. His research interests include deep learning and data mining. (2505817864@qq.com)

Li Liyao received his bachelor's degree from Minnan Normal University in 1994 and master's degree from Fuzhou University in 2009. From 1994 to 2017, he worked at the Fuzhou Branch of Fujian Normal University, during which time he was a visiting scholar at Tsinghua University. Since 2019, he has served as a professor at Fujian Polytechnic Normal University. His research interests include artificial intelligence, network service quality, and cultural digitalization. (leolee@fpnu.edu.cn)

Cairong Ni received her bachelor's degree from Sunshine College in 2022 and has been a teaching assistant at Fujian Normal University of Technology since then. Her research interests include micro-electromechanical systems, big data, and sensors. (3247146792@qq.com)