Vehicle Detection and Distance Estimation Using Improved YOLOv7 Model

ABSTRACT

In this book chapter, we propose a low-cost distance estimation approach to develop more accurate predictions from a 3D perspective for vehicle detection and ranging by using inexpensive monocular cameras. Our distance estimation model integrates YOLOv7 model with an attention module (CBAM) and Transformer as well as extend the prediction vector as the fundamental architecture to improved high-level semantic understanding and enhanced feature extraction ability. This integration significantly improved detection and ranging performance, offering a more suitable and cost-effective solution for distance estimation.

Keywords: Deep learning, YOLOv7, Transformer, Attention module, Vehicle detection and ranging, Scene understanding

INTRODUCTION

Scene understanding serves as a fundamental pillar of driverless technology, as vehicles can only make informed control decisions when they accurately and autonomously perceive the traffic and road scene environment (Gu, et al. 2017). Its ability to provide precise representations and comprehensive understanding of scenes equips it with valuable knowledge of the surroundings, enabling the completion of various tasks in an effective and secure manner (Ignatious, 2023; Liu et al., 2019).

A fundamental component of scene understanding for autonomous vehicles is spatial perception (An, 2021; An & Yan, 2021). spatial perception provides the necessary information for the vehicle to perceive its surroundings accurately, make informed inferences about the scene, ensure safe and reliable autonomous driving (Liu, 2019, Ming, 2021).

From a 3D perspective, spatial perception helps the vehicle estimate the distance and proximity of objects in the scene. This information is crucial for safe navigation, as it allows the vehicle to maintain a safe distance from other vehicles or avoid collisions with obstacles (Sarker et al., 2021; Guo et al., 2021; Zhang et al., 2020; Hu et al., 2020).

Distance estimation can be achieved through various methods, with laser detecting and ranging being a prominent approach for obtaining distance information (Zalevsky et al., 2021). Laser-based distance measurement has gained significant interest in the development of Collision Warning Systems. However, LiDAR technology, though sophisticated, is costly and yields limited results, making it currently suitable only for testing automobiles.

Alternative methods for vehicle detection and distance measurement include ultrasound, infrared, and microwave radar (Aliew, 2022; Özcan et al., 2020). However, each of these approaches has its limitations. Ultrasound and infrared-based distance measurement have certain restrictions,

while microwave radar is sensitive to interference, leading to unreliable detection results. Moreover, these methods often struggle to distinguish between different detection targets.

Hardware-based tools like radar and infrared devices present challenges in terms of costs, integration with imaging devices, and limitations in measurement precision. As a cost-effective solution, discarding expensive distance measurement apparatus was proposed instead of inferring distance information from 2D video footage captured during vehicle detection (Mehtab et al., 2021). This approach offers a potential alternative to the hardware-based methods, but it also comes with its own set of challenges and considerations.

Deep learning-based ranging (distance estimation) holds significant potential for various applications and can be seamlessly integrated with existing approaches to yield superior results. By accurately calibrating the camera internal and external parameters, it becomes possible to determine the distance to the vehicle ahead. This information can then be utilized to provide timely alerts about potential accidents, simulates how objects are projected onto the image plane based on camera parameters or use the visual projection model that based on their appearance and geometric properties to estimate the distance of objects.

In the realm of visual data-based ranging techniques, there are two main branches: Monocular camera-based ranging methods and binocular camera-based ranging methods. Monocular camera ranging relies on initially identifying the target by matching its image with known patterns or features in the scene. This initial identification is typically performed using visual object detection or recognition algorithms, which can locate and classify objects of interest in the image. Once the target object is identified, the distance estimation is based on its apparent size in the image and the knowledge of the real-world size of the object or its category. This method assumes that the physical size of the object is known or can be estimated from prior knowledge.

Within monocular camera ranging, there are several approaches. The circumferential ranging method utilizes a fisheye lens, which can result in more extensive lens distortion (Bremer et al., 2023). However, using a fisheye lens can result in more extensive lens distortion compared to traditional rectilinear lenses. Lens distortion can affect the apparent size and shape of objects in the image, leading to inaccuracies in the perceived dimensions of objects. This impacts the accuracy of distance estimation based on apparent size.

Moreover, fisheye lens distortion is nonlinear and more complex than rectilinear lens distortion. Accurately calibrating the fisheye camera to correct for distortion requires more sophisticated calibration methods and introduces additional computational overhead. Another approach is forward-looking camera ranging, characterized by reduced aberration in the front-view lens helps capture more accurate and undistorted images of the scene in front of the vehicle and the camera being mounted beneath the rearview mirror of the vehicle offers a relatively stable and vibration-free location for the camera, which improves the quality of the captured images (Karimanzira et al., 2021). The third approach is oblique camera ranging, which is distinguished by its larger angle of view (Fukushima, Farzad, & Torras, 2017; Cai et al., 2020). Each of these methods has its unique characteristics and applications in distance estimation using monocular cameras.

In contrast to the distance measurement method used by the forward-facing camera, the circumferential fisheye camera does not rely on mathematical geometry for distance ranging. The reason is that the lens faces downward, resulting in high aberration coefficients, making traditional geometric distance ranging prone to significant errors. Instead, the distance ranging concept for the circumferential fisheye camera is based on a single-strain matrix and affine transformation.

Binocular camera ranging utilizes a pair of cameras to perceive the 3D structure of a scene. This method is inspired by human vision, where our brain uses information from both eyes to estimate depth and perceive the world in three dimensions. In a binocular camera system, the two cameras are positioned side by side, mimicking the separation between human eyes. Each camera captures a slightly different view of the scene due to the horizontal displacement. The images from the two cameras are then employed to compute the disparity, which is horizontal difference between corresponding points in the left and the right images.

Binocular estimation offers several advantages. One notable benefit is that it doesn't require prior recognition of objects, allowing for an unlimited recognition rate. All obstacles can be directly evaluated without the need for pre-existing knowledge. Moreover, binocular estimation doesn't rely on maintaining a sample database, as it operates without the concept of a sample. On the other hand, monocular estimation also comes with its own set of advantages. It is a cost-effective solution, requiring less computational resources, making it more accessible for various applications. Additionally, its relatively simple system design makes it easier to implement and deploy in practical scenes.

The focus of this book chapter is on using deep learning to significantly reduce human workload in distance estimation for vehicle scene understanding. We employ the attention mechanism and Transformer on YOLOv7 as well as extend the prediction vector to estimate the distance between vehicles. Our proposed vehicle ranging model, YOLOv7-CBAM-Transformer, effectively improves the model's understanding of local and global features, thereby enhancing the performance of the original YOLO series models.

We have the related work section, following by the sections methodology and result analysis. Finally, we present the conclusion and future work as well as references.

RELATED WORK

Vehicle detection forms the basis for vehicle ranging, estimating the distance from the vehicle in front is essential for vehicle collision avoidance systems. As a result, an increasing number of articles in the field of computer vision are focusing on the challenges related to vehicle detection and range estimation.

We delve into two areas of literature. In the first area, we explore vehicle detection and range estimation using a binocular camera. Binocular stereo vision involves using two cameras with a known baseline to capture images of the same scene from slightly different viewpoints. Generally, the first step of binocular stereo distance estimation is to make key points or features in the left and right images are identified, and corresponding points are matched between the two images. Then the disparity between matched points is calculated to represent the relative distance of visual

objects in the scene. Finally, using the disparity and known camera parameters to compute the depth information for each pixel is conducted.

A multi-resolution stereovision system was proposed (Chui et al., 2020). The system created image pyramids by down-sampling the captured images to different resolutions. Each level in the pyramid represents a different scale of the scene. The process includes feature extraction, feature matching, disparity calculation, depth estimation, and fusion/refinement steps for each level of the image pyramid. By performing these steps at multiple resolutions, the system can handle visual objects at different distances effectively.

The estimating distance was proffered by identifying corners with high eigenvalues in segmented regions of both images (Alvarado et al., 2022). The model segments the left and right images to identify regions of interest. Image segmentation methods such as thresholding, edge detection, or clustering can be applied to group pixels with similar characteristics into distinct regions. Within each segmented region, a corner detection algorithm was propounded to identify key points or corners. Corner detection algorithms like Harris corner detector or Shi-Tomasi corner detector were accommodated. These algorithms identify points where there are significant variations in intensity in multiple directions, making them suitable for detecting distinctive features. Once the corners are identified, the eigenvalues of the gradient matrix are calculated. The eigenvalues represent the rate of changes of intensity in x-axis and y-axis directions around the corner point. High eigenvalues indicate strong intensity changes, which correspond to well-defined corners. After matching and disparity estimation, triangulation was employed to compute the 3D coordinates of the scene points corresponding to the matched corners in both images. With the 3D coordinates of the matched points, estimate the distance of the objects in the scene from the cameras. The distance can be calculated using simple geometric principles or calibrated camera parameters.

Binocular distance estimation, which relies on using two cameras to capture images of the same scene from slightly different viewpoints, has certain disadvantages when compared to monocular distance estimation, which uses a single camera. Firstly, binocular distance estimation requires the use of two cameras, increasing the hardware complexity and cost compared to monocular systems that only need a single camera. Furthermore, accurate calibration of stereo cameras is crucial for precise distance estimation.

Calibration involves determining the intrinsic and extrinsic parameters of both cameras, any errors in calibration can lead to inaccurate distance measurements. Also, the baseline between the two cameras limits the effective field of view for distance estimation. Visual objects outside this field of view may not be accurately measured by using the stereo vision system. At the same time, in complex scenes, occlusions and disparities between the left and right images can make feature matching and distance estimation challenging. These situations can lead to errors in distance estimation, especially for regions with insufficient texture or ambiguous features. Finally, changes in lighting conditions, reflections, and other environmental factors can affect the performance of binocular distance estimation, making it less robust in certain scenarios.

In contrast, monocular distance estimation, though having its own set of challenges, offers a slew of advantages. Monocular distance estimation requires only one camera, making it a simpler and

more cost-effective solution compared to binocular setups. Since monocular cameras are ubiquitous in a few of devices, it's easier to integrate monocular distance estimation into various applications and platforms. Moreover, monocular cameras can be placed in a variety of positions and orientations, providing more flexibility in system design. At the same time, recent advances in deep learning, especially with monocular distance estimation networks, have improved the accuracy and robustness of monocular distance estimation methods.

Monocular camera-based distance estimation relies on various cues or assumptions. The cues such as perspective, size, and occlusion to estimate distance were observed in the scene (Parker et al., 2022). The distance was estimated by analysing the movement of objects in consecutive frames or using visual odometry techniques (He et al., 2020). Normally, monocular camera-based distance estimation needs to establish correspondences between the extracted features in the image and the corresponding points in the real world (Vijayanarasimhan et al., 2017). This can be achieved by manually identifying the matching points or using known 3D reference points. Through using the camera calibration parameters and the matched feature correspondences, triangulation was employed to obtain 3D coordinates of real-world points. Triangulation was utilized to calculate the intersection of rays originating from the camera center and passing through the matched feature points. Since the initial 3D coordinates are only up to an unknown scale factor, the known dimensions are needed to estimate the scale and convert the 3D coordinates to actual world coordinates. Once the actual world coordinates of the structures, objects, or road segments are obtained, the distances between points of interest in the scene can be computed by measuring the Euclidean distance between their corresponding 3D coordinates.

There has monocular distance estimation using inverse perspective mapping (IPM) from a bird's-eye view. This transformation allows to estimate distances directly in the transformed view, which simplifies distance estimation. A perspective transformation (IPM) was deployed to map the image from the view perspective to a bird's-eye view. In the bird's-eye view, parallel lines become parallel and perpendicular to the ground, simplifying distance estimation. A mapping between the bird's-eye view and the real-world coordinates is created. This mapping relates the pixel coordinates in the transformed view to the corresponding real-world 3D points (Vakili et al., 2020).

The integration of attention mechanisms was employed to enhance the accuracy of distance estimation models. An innovative strategy has emerged that integrates global relative constraints to promote consistent vehicle state estimations. This methodology emphasizes on the significance of capturing both contextual and spatial details during the estimation process. The architecture of MSANet, the proposed framework, was elaborated, encompassing distinct streams for motion, context-awareness, and spatial information extraction from input data. To achieve enhanced estimation accuracy, the multi-stream attention fusion (MSAF) block was introduced as a means to effectively amalgamate these distinct features (Huang, Huang & Hsu, 2021).

Furthermore, Transformer and attention modules have been integrated. An advanced system rooted in deep learning was designed to autonomously identify physical distancing through the analysis of surveillance footage from security cameras. In this approach, TH-YOLOv5 was adopted for the purpose of object detection and classification, while DeepSort was employed to track individuals detected within bounding boxes outlined in the video material.

An innovative facet of this methodology involves the incorporation of Transformer Heads (TH) into the TH-YOLOv5 framework. This addition capitalizes on the self-attention mechanism, thus augmenting the model's predictive capabilities. Furthermore, the Convolutional Block Attention Model (CBAM) was introduced to pinpoint regions of interest within densely populated scenes. To achieve this, specific convolutional and bottleneck blocks were replaced with transformer encoder blocks, drawing inspiration from the architecture of vision transformers. This adaptation enables more comprehensive acquisition of global and contextual information, proving especially advantageous in intricate scenarios involving occlusions. These transformer encoder blocks are seamlessly integrated into the head segment of the backbone, enhancing feature representation and ultimately contributing to heightened object detection performance.

Moreover, the incorporation of CBAM module aided the TH-YOLOv5 model in directing its attention towards pertinent target elements present in images captured by using CCTV cameras. The system estimated the depth between the camera and objects by utilizing coordinate transformations and camera intrinsics. The pairwise L2 normalization was harnessed to calculate the distance between tracked individuals, and a violation index is computed to identify breaches of physical distance rules (Junayed & Islam, 2022).

In summary, In order to highlight our contributions of the proposed work and compare it with the achieved results in the literature (Liu, et al. 2019, Liu, et al. 2022), we would like to emphasize on that we adopt the attention mechanism and Transformer on YOLOv7 to estimate the distance between vehicles. Our proposed vehicle ranging model, YOLOv7-CBAM-Transformer, effectively improves the model understanding of local and global features, thereby enhancing the performance of the original YOLO series models.

METHODOLOGY

In our distance estimation model, we adopt YOLOv7 as the foundational architecture. Additionally, we integrate Swin Transformer and Convolutional Block Attention module into the model to enhance the feature extraction capabilities further (Woo et al., 2019; Chienyao, Alexey & Mark, 2022; Liu et al., 2021).

The design of YOLOv7 aims to address two specific challenges. Firstly, it introduces the concept of gradient propagation routes, which facilitates a structured model re-referencing approach. This approach enables the analysis of structural re-referencing techniques that are pertinent to each network layer.

Training models with multiple output layers using a dynamic label assignment poses additional challenges, particularly concerning the assignment of dynamic targets to the outputs of different branches. To address this challenge, a novel approach was proposed called the coarse-to-fine guided label assignment technique for labeling assignment. This method offers a solution to overcome the issue of assigning dynamic targets to the outputs of various branches.

Our model, as illustrated in Fig.1, incorporates the Convolutional Block Attention Module (CBAM) to enhance the feature extraction process, thereby avoiding alterations to the original feature extraction. Moreover, the inclusion of the Transformer enhances the model capacity to

comprehend global semantics, allowing YOLOv7, which primarily emphasizes on local information processing, to achieve a more comprehensive understanding of traffic scenes. The Convolutional Block Attention Module encompassing the Channel Attention (CAM) and the Spatial Attention Module (SAM) (Woo, Park, Lee & Kweon, 2019). The CAM computes attention maps by analysing the interdependencies among channels, determining which channels contain significant information for a given context. The channel attention mechanism allows the network to emphasize important channels while reducing the influence of less relevant ones.

On the other hand, the SAM computes attention maps to highlight spatial regions containing relevant information. This mechanism enables the network to adaptively focus on specific parts of the image while downplaying background or less informative areas. The CAM focuses on capturing channel-wise relationships within the feature maps. It consists of an MLP followed by an element-wise summation and a sigmoid activation function. The MLP processes the input feature map and computes channel-wise attention weights that highlight informative channels and suppress less relevant ones. The attention weights obtained from the MLP are scaled by using the sigmoid activation function to ensure they lie within the range of 0 to 1. The scaled attention weights are then applied element-wise to the original feature map to enhance informative channels,

$$\mathrm{M_C(F)} = \sigma\left(\mathrm{MLP}\big(\mathrm{AvgPool(F)}\big) + \mathrm{MLP}\big(\mathrm{MaxPool(F)}\big)\right) = \sigma(\mathrm{W_1}\left(\mathrm{W_0}\big(\mathrm{F^c_{avg}}\big)\right) + \mathrm{W_1}\big(\mathrm{W_0}(\mathrm{F^c_{max}})\big)) \quad (1)$$

where σ represents the sigmoid function, while W_0 and W_1 are the shared MLP weights for both inputs, $W_0 \in \mathbb{R}^{\mathrm{C/r\times C}}$, and $W_1 \in \mathbb{R}^{\mathrm{C\times C/r}}$. The ReLU activation function is applied following these weights.

The SAM analyzes spatial relationships within the feature maps to highlight important spatial regions. It involves two operations: Max pooling and convolution. Max pooling captures global context by summarizing the most significant spatial information within each channel. Convolution captures local context by processing the feature map with a convolutional filter to capture spatial dependencies. The outputs of max pooling and convolution are combined by using an element-wise summation. The combined outputs undergo a sigmoid activation to produce spatial attention weights. The spatial attention weights are applied to the feature map to highlight relevant spatial regions (Li et al., 2023),

$$\mathrm{M_s(F)} = \sigma\big(\mathrm{f^{7\times7}([AvgPool(F); MaxPool(F)])}\big) = \sigma(\mathrm{f^{7\times7}([F^s_{avg}; F^s_{max}])}) \quad (2)$$

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

CBAM has demonstrated its effectiveness in improving the discriminative power of CNNs and boosting their accuracy on various benchmarks and datasets. Its ability to adaptively emphasize on salient features while suppressing irrelevant ones helps networks achieve better generalization and robustness (Pan, 2018; Pan, 2020; Pan, 2021), making CBAM a valuable tool for advancing the capabilities of deep learning models in the field of computer vision.

Nonetheless, YOLOv7 network holds a significant advantage in extracting foundational features and visual structures. These low-level features encompass essential points, lines, and fundamental image elements at the patch level. These fundamental features exhibit distinct geometric characteristics and often emphasize consistency or covariance under transformations such as translation and rotation. Once the fundamental visual components are identified, the emphasis shifts towards comprehending the advanced visual meaning. This centers on grasping the interconnections among these components, shaping them into objects, and perceiving how the spatial arrangement of objects generates a scene.

Presently, the Transformer model is widely regarded as proficient and efficient in managing the intricate relationships among these components. As a result, we remove the last ELAN in the YOLOv7 backbone and ELANs in the neck and instead integrate the Swin Transformer encoder. By implementing this operation, we accentuate the benefits of the self-attention mechanism while simultaneously reducing computational overhead. Moreover, the introduction of Transformer in the neck allows for capturing correlations and significance between different regions, enhancing the model ability to adapt to targets of varying sizes (Zhang, 2023).

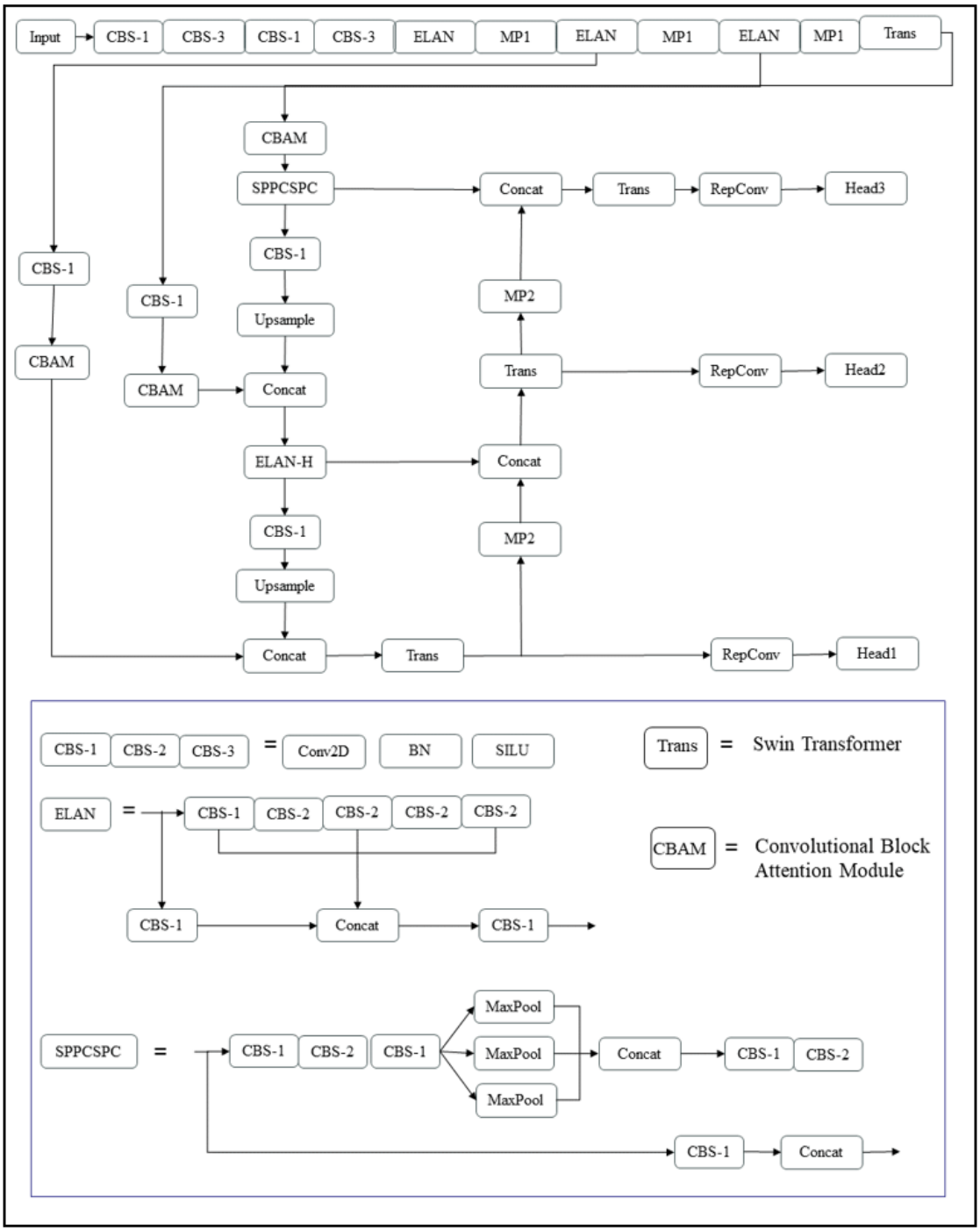

Fig. 1 *The architecture of YOLOv7-CBAM-Transformer*

As we implement YOLOv7-CBAM-Transformer, our goal is not only to detect the position of the vehicle but also to estimate the distance between the vehicle in front and the current position. To achieve this, we have extended the prediction vector to incorporate distance estimation information.

The original prediction vector contains bounding box anchor coordinates **A**(x, y, w, h) and category confidence **C** (c_1, c_2). In order to make the model realize the ranging function, we add the distance element **D** (d) to the prediction vector. The extended prediction vector is shown in the Fig. 2.

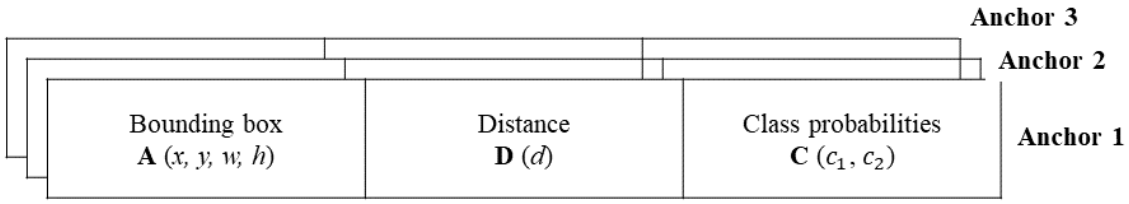

Fig. 2 *The extended prediction vector for distanace estimation*

The distance loss is defined as

$$l_{distance}(i,j) = \omega(d'_{i,j} - d_{i,j})^2 = \omega \sum_{k=0}^{c} C_{i,j,k}(d'_{i,j,k} - d_{i,j,k})^2 \quad (3)$$

where $C_{i,j,k}$ is k-th class probability in (i, j)-th cell. The weighting constant ω is introduced to balance the importance of the distance loss with other losses, preventing it from dominating the overall training process. In our experiment, we set ω to a value of 1×10^2.

RESULT ANALYSIS

In this book chapter, we present a novel vehicle detection and distance estimation model for low-cost monocular cameras, enhanced with an attention module and Transformer, utilizing deep learning techniques. The experimental setup involved using PYTHON 2.7, an RTX5000 GPU, and 32GB RAM. The example data samples in the experiments were from the KITTI dataset. The KITTI dataset comprises both intrinsic and extrinsic characteristics of the in-car camera, along with the coordinates, width, and height of the detection boxes. For the development of our deep learning model, we randomly selected 4,000 samples and divided them into a 7:3 ratio for training and testing. The results presented in this section correspond to state-of-the-art approaches. Fig. 3 illustrates the satisfactory vehicle recognition and distance estimation performance achieved by our modified YOLOv7 (YOLOv7-CBAM-Transformer) with the extended prediction vector.

To train our YOLOv7-CBAM-Transformer, we set the following parameters: *epochs* as 5,000, *batch size* as 1.0, and *learning rate* as 0.01. Fig. 4 illustrates the network training process, shown that validation loss decreases steadily between 0 and 1,000 iterations. After reached 1,000 epochs, the loss curves stabilize at 0.082.

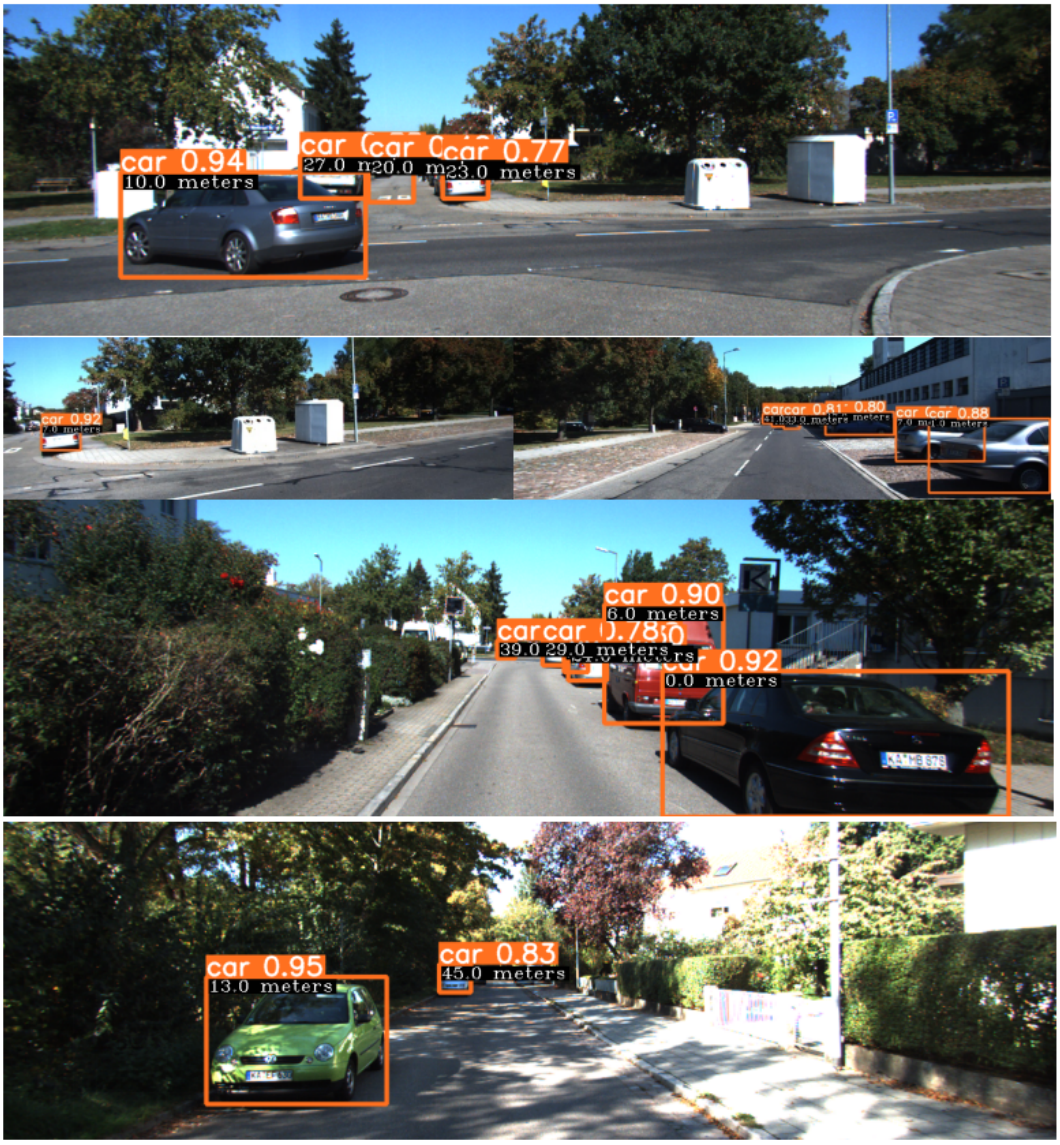

Fig. 3 *The example of vehicle detection and distance estimation using YOLOv7-CBAM-Transformer*

Table 1 shows a quantitative comparison of the KITTI-constructed dataset for each of the evaluation measures listed. YOLO models and the transformer model are evaluated in this comparison. The results indicate that YOLOv7 outperforms all previous YOLO models, including the transformer. Moreover, our YOLOv7-CBAM-Transformer, which incorporates the convolutional block attention module, outperforms the original YOLOv7. The combination of the convolutional block attention module with YOLOv7 demonstrates significantly improved results. Notably, the addition of the Swin Transformer leads to a reduction of 0.382 in RMSE compared to the previous model. Overall, our YOLOv7-CBAM-Transformer achieves a total reduction of 0.456 in RMSE compared to the original YOLOv7 model.

Additionally, the distances were divided into three categories: 0-10m, 10-20m, and >20m. In Table 2, we obtain the average RMSE for each group. Our YOLOv7-CBAM-Transformer outperforms

the original YOLOv7 in 0-10*m* and 10-20*m* distance categories. In summary, the data presented in Table 1 and Table 2 shows that the YOLOv7-CBAM-Transformer model is much effective in handling object detection and distance estimation tasks. Moreover, the model with Transformer is reduced by at least 0.2 in the RMSE of each distance category compared with YOLOv7-CBAM. Compared with the original YOLOv7, it is reduced by at least 0.303 in distance category of 0-10*m* and 10-20*m*.

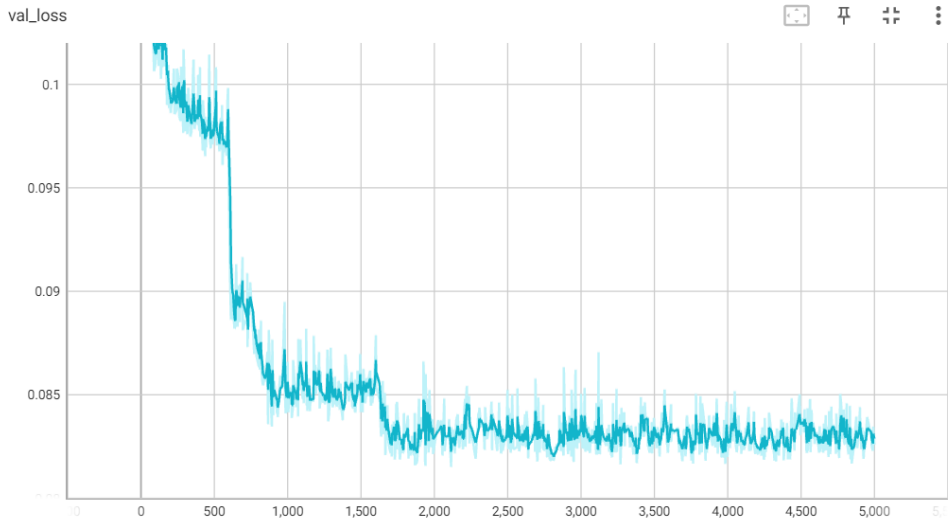

Fig. 4 *The diagram of training process of YOLOv7-CBAM. The blue curve indicates the validation loss.*

Table 1 Comparative Analysis of Multiple Deep Neural Networks

Models	RMSE
YOLOv5	4.121
YOLOv6	4.483
YOLOv7	4.157
YOLOv7-CBAM	4.083
YOLOv7-CBAM-Transformer	3.701
Swin Transformer	4.776

Table 2 Average RMSE of different neural networks in different distance

Models	0-10m	10-20m	>20m
YOLOv7	4.502	4.357	3.612
YOLOv7-CBAM	4.499	3.667	4.083
YOLOv7-CBAM-Transformer	4.199	3.021	3.883

Table 3 Comparing our model to other state of the art model of distance estimation

Model	RMSE
GC-ASPP-YOLOv3-D (Lian et al., 2022)	3.985
Ours (YOLOv7-CBAM-Transformer)	3.701

In the context of identical dataset, we conducted a comparative analysis between our model and another estimation model GC-ASPP-YOLOv3-D in Table 3 (Lian et al., 2022). The results highlight the distinct advantages of our model, particularly in terms of RMSE, where it outperforms GC-ASPP-YOLOv3-D.

In summary, in addition to YOLOv7 being more suitable for vehicle distance estimation in KITTI dataset than YOLOv5, YOLOv6 and original Swin Transformer, adding CBAM can successfully further reduce the estimation error of the model. Moreover, we found that YOLOv7 with CBAM is much suitable for distance estimation within 20 meters, while the original YOLOv7 is prominent when estimating the distance of vehicles beyond 20m. By further improvement, our proposed YOLOv7-CBAM-Transformer produced the better results even compare with YOLOv7-CBAM.

CONCLUSION

Our primary focus was on developing an advanced deep learning-based model tailored for low-cost monocular cameras, with the aim of optimizing hardware costs (Gowdra, et al., 2021). By accurately detecting vehicles and applying extended distance estimation vector, our model obtains the bounding box coordinates, enabling precise distance calculations. Through experimentation, we discovered that combining YOLOv7-CBAM-Transformer with the extended distance estimation vector yielded the best results, achieved a remarkable 0.456 improvement in RMSE compared to the original YOLOv7 model. In our analysis of the results, we observed that the YOLOv7-CBAM-Transformer outperforms the original YOLOv7-CBAM model in distance measurement. It exhibited enhanced accuracy and reduced occurrences of false detections and missed detections. This highlights the significant impact of the Transformer, which contributed to improve the feature understanding and its adaptability to diverse distances of vehicles in the

scenes. In future, we would like to explore how to use generative pre-trained transformer models to measure the distance between vehicles (Yam, 2019; Yan, 2023).

REFERENCES

Alexey, B., ChienYao, W., & Mark, L. (2020). YOLOv4: Optimal speed and accuracy of object detection. Image and Video Processing.

Alfred Daniel, J. et al. (2023). Fully convolutional neural networks for LIDAR–camera fusion for pedestrian detection in autonomous vehicle. Multimedia Tools and Applications, pp.1-24.

Aliew, F. (2022). An approach for precise distance measuring using ultrasonic sensors. Engineering Proceedings, 24(1), pp.8.

Alvarado, S. T., Borja, M. G. B., & Torres, K. B. (2022). Object distance estimation from a binocular vision system for robotic applications using artificial neural networks. Control, Mechatronics and Automation (ICCMA), pp. 19-23.

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. *ACM Transactions on Multimedia Computing, Communications and Applications.*

Bremer, J., Maj, M., Nordbø, Ø., & Kommisrud, E. (2023). Deep learning–based automated measurements of the scrotal circumference of Norwegian red bulls from 3D images. Smart Agricultural Technology, 3, pp.100133.

Cai, Y., Ding, Y., Zhang, H., Xiu, J., & Liu, Z. (2020). Geo-location algorithm for building targets in oblique remote sensing images based on deep learning and height estimation. Remote Sensing, 12(15), pp.2427.

Chienyao, W., Alexey, B., Mark, L. (2022). YOLOv7: Trainable bag-of-freebies sets newstate-of-the-art for real-time object detectors. Computer Vision and Pattern Recognition, arXiv:2207.02696

Fukushima, H., Farzad, D., Babette & Torras, C. (2017). Scene Understanding Using Deep Learning. Academic Press, pp. 373-382.

Gowdra, N., Sinha, R., MacDonell, S., Yan, W. (2021) Maximum Categorical Cross Entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in Convolutional Neural Networks (CNNs) by reducing overfitting. *Pattern Recognition.*

Gowdra, N. (2021) *Entropy-Based Optimization Strategies for Convolutional Neural Networks.* PhD Thesis, Auckland University of Technology, New Zealand.

Gu, Q., Yang, J., Kong, L., Yan, W., Klette, R. (2017) Embedded and real-time vehicle detection system for challenging on-road scenes. *Optical Engineering*, 56 (6), 063102.

Gu, Q., Yang, J., Yan, W., Klette, R. (2017) Integrated multi-scale event verification in an augmented foreground motion space. Pacific-Rim Symposium on Image and Video Technology (pp.488-500)

Gu, Q., Yang, J., Yan, W., Li, Y., Klette, R. (2017) Local Fast R-CNN flow for object-centric event recognition in complex traffic scenes. Pacific-Rim Symposium on Image and Video Technology (pp.439-452)

Guo, J., Wang, J., Wang, H., Xiao, B., He, Z., & Li, L. (2023). Research on road scene understanding of autonomous vehicles based on multi-task learning. Sensors, 23(13), pp.6238.

He, Z., Yang, Q., Zhao, X., Zhang, S., & Tan, J. (2020). Spatiotemporal visual odometry using ground plane in dynamic indoor environment. Optik, pp.165165.

Huang, K. C., Huang, Y. K., & Hsu, W. H. (2021). Multi-stream attention learning for monocular vehicle velocity and inter-vehicle distance estimation. arXiv preprint arXiv:2110.11608.

Ignatious, H. A et al., (2023). Analyzing factors influencing situation awareness in autonomous vehicles—A survey. Sensors, 23(8), pp.4075.

Junayed, M. S., & Islam, M. B. (2022). Automated physical distance estimation and crowd monitoring through surveillance video. SN Computer Science, 4(1), pp.67.

Karimanzira, D., Pfützenreuter, T., & Renkewitz, H. (2021). Deep learning for long and short range object detection in underwater environment. Adv Robot Automn 5(1), pp.1-10

Li, H., Tan, Y., Miao, J., Liang, P., Gong, J., He, H., ... & Wu, D. (2023). Attention-based and micro designed EfficientNetB2 for diagnosis of Alzheimer's disease. Biomedical Signal Processing and Control, 82, pp.104571.

Lian, G., Wang, Y., Qin, H., & Chen, G. (2022). Towards unified on-road object detection and depth estimation from a single image. Machine Learning and Cybernetics, pp.1-11.

Liu, X. (2019). Vehicle-related Scene Understanding Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand

Liu, X., & Nguyen, M., Yan, W. (2019). Vehicle-related scene understanding using deep learn. ACPR, pp. 61-73

Liu, X., Yan, W., & Kasabov, N. (2020). Vehicle-related scene segmentation using CapsNets. IEEE IVCNZ, pp. 1-6.

Liu, X., & Yan, W. (2021). Traffic-light sign recognition using Capsule network. Springer Multimedia Tools and Applications, pp. 15161-15171.

Liu, X. &Yan, W. (2022). Depth estimation of traffic scenes from image sequence using deep learning, PSIVT.

Liu, X. & Yan, W. (2022) Vehicle-related distance estimation using customized YOLOv7. *International Conference on Image and Vision Computing New Zealand* (IVCNZ)

Liu, X. & Yan, W. Kasabov, N. (2022) Moving vehicle tracking and scene understanding: A hybrid approach. Multimedia Tools and Applications.

Liu, Z, et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. Computer vision, pp.37-49

Masoumian, A., Rashwan, H. A., Cristiano, J., Asif, M. S., & Puig, D. (2022). Monocular depth estimation using deep learning: A review. Sensors, 22(14), pp.5353.

Mehtab, S., Yan, W. (2021) FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. *International Conference on Control and Computer Vision*.

Mehtab, S., Yan, W. (2022) Flexible neural network for fast and accurate road scene perception. *Multimedia Tools and Applications.*

Mehtab, S. Yan, W., Narayanan, A. (2022) 3D vehicle detection using cheap LiDAR and camera sensors. *International Conference on Image and Vision Computing New Zealand.*

Mehtab, S. (2022) *Deep Neural Networks for Road Scene Perception in Autonomous Vehicles Using LiDARs and Vision Sensors.* PhD Thesis, Auckland University of Technology, New Zealand.

Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. Neurocomputing, 438, pp.14-33.

Ming, Y., Li, Y., Zhang, Z., Yan, W. (2021) A survey of path planning algorithms for autonomous vehicles. *International Journal of Commercial Vehicles*.

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. *Multimedia Tools and Applications*, 79 (27-28), 19925-19944.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. *IEEE Transactions on Image Processing.*

Özcan, M., Aliew, F., & Görgün, H. (2020). Accurate and precise distance estimation for noisy IR sensor readings contaminated by outliers. Measurement, 156, pp.107633.

Parker, P. R., Abe, E. T., Beatie, N. T., Leonard, E. S., Martins, D. M., Sharp, S. L., ... & Niell, C. M. (2022). Distance estimation from monocular cues in an ethological visuomotor task. Elife, pp.74708.

Vakili, E., et al. (2020). Single-camera vehicle speed measurement using the geometry of the imaging system. Mult. Tools Apps. 79, pp.19307–19327

Vijayanarasimhan, S., et al. (2017). Sfm-net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2019). CBAM: Convolutional blockattention module. Computer Vision, pp. 3-19

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer

Yan, W. (2023) Computational Methods for Deep Learning – Theory, Algorithms, and Implementations (2nd Edition). Springer

Zalevsky, Z. et al., (2021). Light detection and ranging (Lidar): Introduction. JOSA A, 38(11), pp.LID1-LID2.

Zhang, D. (2023). STA-YOLOv7: Swin-transformer-enabled YOLOv7 for road damage detection. Computer Science and Application. 13. pp.1157-1165.