# Vehicle Detection Algorithm Based on Fisheye Camera in Parking Environment

Liao Wang and Qiu Fang
*Xiamen University of Technology, Xiamen, China*

Keywords: Deep Learning, YOLOv5, Fisheye Image, Distortion, Object Detection.

Abstract: To address the issue of missing and wrong detection of fisheye images by existing target detection algorithms, a targeted dataset is constructed, and an improved model is proposed using YOLOv5 as a reference. Firstly, to facilitate better adapt to the dimensions of the custom dataset, the K-means++ algorithm was utilized for anchor box clustering. Secondly, a deformable convolutional network was brought in to substitute some convolution layers in the original network, so that the network can adaptively extract distorted image feature points, and the integration of coordinate attention mechanisms enhances the expression of semantic and positional information of the feature points of interest. Furthermore, Slim-Neck is designed to replace the original Neck based on GSConv convolution, resulting in reduced model parameters and enhanced algorithmic precision rate. Lastly, redesigning the detector's loss function by incorporating EIoU Loss and Focal Loss. The results demonstrate that precision rate, recall rate and mean precision are improved by 2.31%, 4.41% and 3.50% respectively compared with YOLOv5 algorithm.

## 1 INTRODUCTION

Autonomous valet parking (AVP) plays a pivotal role in self-driving, the implementation of this technology relies on the perception of the environment by sensors such as fisheye cameras and ultrasonic radar. Among them, fisheye cameras have unique advantages in detection due to their wide field of view and distortion characteristics(Lee M, 2019). Due to the large amount of aberrations in their acquired images, it becomes a challenging problem to detect vehicles quickly and accurately in fisheye images. In recent years, many scholars have used neural networks to accomplish target detection. Deep learning-based object detection algorithms can be categorized into two-stage object detection algorithms, exemplified by R-CNN (Girshick R, 2014) and Fast R-CNN (Yang, 2020) and single-stage object detection algorithms, exemplified by YOLO (Redmon, 2018) and SSD (Chen X, 2019), according to the presence or absence of the displayed regions suggested by the algorithms (Jun Jiang, 2021). In terms of target detection based on fisheye images, the literature (Wei, 2022) concluded that the main problems affecting the detection accuracy of fisheye images are object rotation and spatial distortion. The authors proposed a rotation mask deformable convolutional network architecture to improve the capacity for learning and computational efficiency of the convolutional kernel for rotating object features in fisheye images, but its model complexity is high. The literature (Fremont, 2016) divides the image into 7 regions and generates 4 sets of training samples for 4 different distortion levels and imaging models. The method has good performance in pedestrian detection task in fisheye images, but still has the problem of missed and wrong detection. Hence, this paper presents a deep learning-based algorithm for object detection in fisheye images.

The primary tasks are as follows: (1) constructing a fisheye dataset based on the parking environment; (2) introducing the deformable convolution into the feature extraction network, we propose the DCNC3 module as a replacement for original C3 module, and incorporating the coordinate attention mechanism; (3) designing the GSCSP structure based on the new hybrid convolution GSConv to optimize the model performance.

Wang, L. and Fang, Q.
Vehicle Detection Algorithm Based on Fisheye Camera in Parking Environment.
DOI: 10.5220/0012272600003807
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 5-10
ISBN: 978-989-758-677-4
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

## 2 DATASET CONSTRUCTION

The only publicly available fisheye dataset for the autonomous driving domain is WoodScape from Valeo (Yogamani, 2019). Given that the scenarios of this dataset are mostly foreign and parking scenarios are few. Therefore, we constructed our own fisheye dataset based on the parking environment. The fisheye cameras were mounted above the air intake grille, on the front left and right door panels, and above the rear license plate (Figure 1). To ensure the diversity of the dataset, images of underground and open-air parking lots were collected under different weather and at different time periods according to occupying parking spaces in Xiamen and Shanghai. The dataset is divided into training, testing, and validation sets with an 8:1:1 distribution. 8400 images were obtained by manually labeling the data with LabelImg, a data labeling software. An illustration of the dataset is depicted in Figure 2.

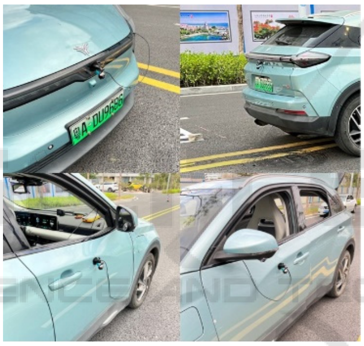

Figure 1: Fisheye camera installation position.

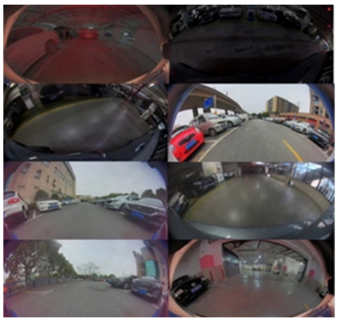

Figure 2: Example of dataset.

There are more small targets as well as distorted shapes in this dataset with large variations in pixel sizes. Therefore, using the k-means++ algorithm to perform clustering on frames containing vehicle labels. to obtain the most suitable a priori frame size for this dataset, and the a priori frame size after clustering is shown in Table 1.

Table 1: K-means++ clustering a priori box.

| Feature Map | Receptive Field | Anchor |
|---|---|---|
| 13*13 | Big | (113,67)(93,99)(138,109) |
| 26*26 | Middle | (41,38) (76,41) (66,70) |
| 52*52 | Small | (16,12) (27,20) (48,24) |

## 3 IMPROVED NETWORK ALGORITHM BASED ON YOLOv5

YOLOv5 adopts Anchor-based detection method, which belongs to single-stage target detection method. Compared with the previous versions, YOLOv5 offers enhanced speed and improved accuracy, and is one of the leading target detection algorithms in the industry. The algorithm's core concept is to partition the image into grids, each responsible for predicting object type and location. It filters target boxes based on IoU (Intersection over Union) values between predicted and actual bounding boxes, ultimately outputting class labels and positional coordinates for predicted bounding boxes. YOLOv5 is structured around four key elements: input, backbone network, neck feature fusion network, and head module. Aiming at the problems in fisheye image detection, this paper constructs a vehicle detection model utilizing the fisheye camera based on the YOLOv5s model. Its architecture is illustrated in Figure 3.

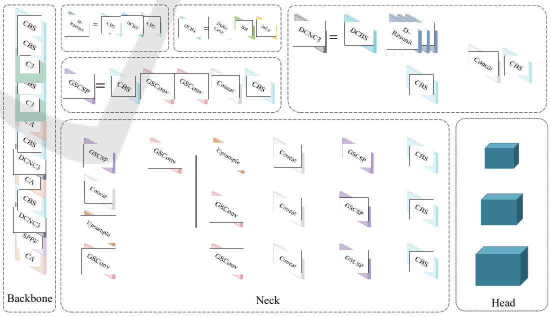

Figure 3: The network architecture of the model.

### 3.1 DCBS Module Based on Deformable Convolution

A deformable convolution module (DCBS) is introduced into the feature extraction framework. As show in Figure 4 this module is composed of deformable convolution layer, BN layer and a SiLu activation function layer. Some C3 modules in the feature extraction network are substituted with DCNC3 modules. The DCNC3 structure consists of a conventional convolution module (CBS), a deformable convolution module (DCBS), and N D-Resunit residuals (only 3×3 convolution layers are replaced by deformable convolution).Deformable convolution can alter the shape and size of the receptive field based on the distorted image.

Traditional convolution uses fixed-size convolution kernel sampling and weighting for feature graphs. The calculation formula is outlined below:

$$y(p_0)=\sum_{p_n\in R}\omega(p_n)\mathrm{g}x(p_0+p_n) \qquad (1)$$

On the basis of formula 1, the offset {ΔPn | n=1,2,3...,N},N=|R| is extended for all points in grid R. Then $y(p_0)$ in the output feature diagram is adjusted as follows :

$$y(p_0)=\sum_{p_n\in R}\omega(p_n)\mathrm{g}x(p_0+p_n+\mathrm{V}p_n) \qquad (2)$$

Then add a weighting factor for the sampling points$\Delta m_n \in [0,1]$ The formula is as follows:

$$y(p_0)=\sum_{p_n\in R}\omega(p_n)\mathrm{g}x(p_0+p_n+\mathrm{V}p_n)\times\mathrm{V}m_n \qquad (3)$$

Finally, the offset pixels are obtained by using bilinear interpolation as illustrated below:

$$X(p)=\sum_{q}G(q,p)\mathrm{g}X(q) \qquad (4)$$

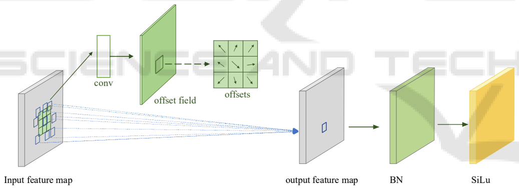

Figure 4: Structure of the DCBS module.

## 3.2 Integration of Coordinate Attention Mechanism

To facilitate the network in autonomously directing its attention towards distorted targets, the coordinate attention(CA) mechanism is incorporated in this paper (Hou, 2021). The module's structural configuration is depicted in Figure 5. The CA attention module first performs the average pooling operation along the vertical and horizontal orientations of the input feature map, and completes the cutting of the number of channels through Concat and 1*1 convolution operation. Then the spatial information is encoded through the BN layer and the non-linear layer, and the normalization weighting process is carried out at the same time.

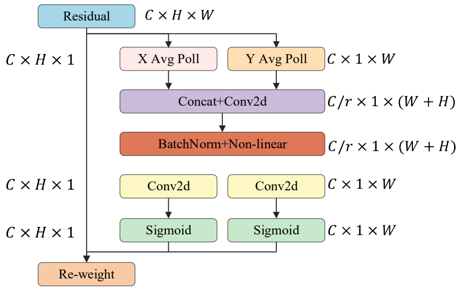

Figure 5: Architecture of coordinate attention mechanism.

## 3.3 GSCSP Module

In this paper, a hybrid convolution method GSConv is introduced by means of standard convolution(SC), depth separable convolution(DSC) (Chollet, Xception, 2017) and shuffle combination and its structure is shown in Figure 6. As shown in Figure 7, based on convolutional GSConv, the GSbottleneck structure is introduced, and Gsbottleneck replaces the Bottleneck structure of the C3 module within the feature integration network with GSbottleneck. The Gsbottleneck structure replaces the 1×1 convolution layer with GSConv and adds a new skip connection. This improvement can reduce the number of network parameters, thereby improving the computational efficiency of the network and mitigating the risk of overfitting. In addition, the addition of new jump connections makes the two branches do not share the same weight, and the information of each channel experiences a unique propagation path, so the correlation and difference of channel information are significantly enhanced, which not only improves the transmission efficiency of information, but also accelerates the convergence speed of training. Remove the C3 module from the original Neck and add the GSCSP module. Finally, the GSConv, GSbottleneck and GSCSP modules can be combined to form a Slim-Neck network.

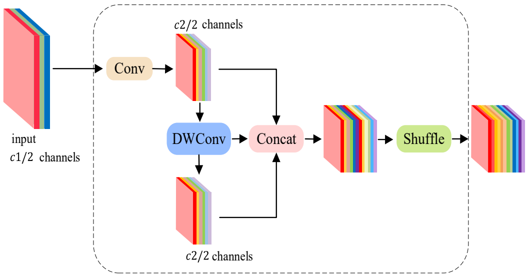

Figure 6: Structure of GSConv module.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

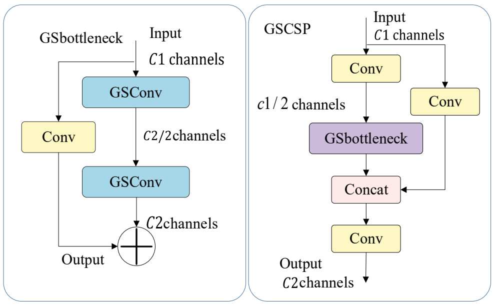

Figure 7: Structure of Bottleneck and GS bottleneck.

# 4 EXPERIMENTS AND ANALYSIS OF RESULTS

## 4.1 The Experimental Setup and Assessment Metrics

The experiments in this paper are carried out on Ubuntu18.04 operating system, the hardware environment is Intel(R) Core(TM) i7-9700 CPU@2.60GHz, the memory is 32G, the GPU is GeForce RTX 3090, the video memory is 24G. The software environment is python3.8, and the deep learning framework is Pytorch 1.11.0. The image size resolution is 640×640, the training batch size is 32, and the network model parameters are learned and updated by SGD. The initial learning rate is 0.01, the learning rate decline parameter is 0.0001, the momentum is 0.937, the weight decay coefficient is 0.0005, and the training times (epochs) of all samples in the training set is 150. In the experiment, the mean Average Precision (mAP), Precision and Recall were used as evaluation indicators to evaluate the model. The formula for calculating the index is as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

$$AP = \int_0^1 precision(t)dt \tag{7}$$

$$mAP = \frac{\sum_{n=1}^{N} AP_n}{N} \tag{8}$$

In this context, TP denotes the quantity of positive samples correctly predicted, FP represents the number of negative samples mistakenly predicted as positive, and the count of positive samples erroneously classified as negative is denoted by FN.

## 4.2 Analysis of Experimental Results

1) Algorithm Performance Comparison Experiment

For the purpose of validating the advancement of the algorithm in this paper, YOLOv3, YOLOv4, YOLOv5, SSD, Faster-RCNN, Retinanets and the improved algorithm are selected to compare the objective evaluation indicators in the data set constructed in this paper. The comparison results of their detection performance are shown in Table II. Table 2 shows that the precision, recall and mean average precision of the proposed algorithm are greatly improved compared with other algorithms. Compared with the original YOLOv5 algorithm, the precision, recall and mean average precision are increased by 2.31%, 4.41% and 3.50%, respectively. In conclusion, the improved algorithm in this paper has better detection performance.

Table 2: Different model detection performance comparison.

| Model | mAP@0.5 | Precision | Recall |
|---|---|---|---|
| SSD | 75.25 | 80.0 | 53.53 |
| Faster-RCNN | 74.79 | 78.63 | 64.61 |
| Retinanets | 67.43 | 87.32 | 55.21 |
| YOLOV3 | 88.35 | 89.24 | 79.93 |
| YOLOV4 | 82.56 | 91.28 | 73.5 |
| YOLOV5 | 91.40 | 89.39 | 84.99 |
| Ours | 94.90 | 91.70 | 89.40 |

2) Comparison of the Model Before and After Improvement

To assess the effectiveness of the aforementioned improvement strategies on the missed detection and error detection problem, the ablation experiment is carried out. Table 3 shows the average precision mean, precision and recall rate of different improvement strategies. Table 2 illustrates the consequences of integrating deformable convolutions in the reconstruction of the backbone network, the average precision mean, precision and recall rate are increased by 2.36%, 1.12% and 1.36%, respectively. After the introduction of Slim-Neck, although the mean average precision is increased by 1.05%, the precision is only increased by 0.28%, but the recall rate is increased by 2.22%. It is evident that this improvement effectively decreases the occurrence of missed detections and false detection rate. The improved experiments show that DCNC3, Slim-Neck and CA have the potential to boost detection accuracy to a certain degree.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 3: Experimental results of the improved model.

| Baseline algorithms | Improvement strategy | | | mAP@50 | Precisi on | Recall |
|---|---|---|---|---|---|---|
| | DCNC 3 | CA | GSCSP | | | |
| YOLOv5 | | | | 91.40 | 89.39 | 84.99 |
| | √ | | | 93.76 | 90.51 | 86.35 |
| | √ | √ | | 93.85 | 91.70 | 87.18 |
| | √ | √ | √ | 94.90 | 91.98 | 89.40 |

3) Assessment of model performance on The WoodScape Dataset

To further reflect the advancement of the proposed algorithm, this paper verifies it on WoodScape, the public dataset of Valeo autonomous driving fisheye. The improved algorithm and the original YOLOv5 algorithm are verified by using this data set. Refer to Table IV for the findings of the experiments. A comparative analysis was carried out in relation to the original YOLOv5 algorithm, the mean average precision, precision and recall of the proposed algorithm are increased by 0.93%, 0.61% and 0.96%, respectively. Since the images in the dataset constructed in this paper are all domestic parking scenes, the types of detected vehicles are not as rich as those in the WoodScape dataset. Therefore, although the algorithm's precision in detecting objects on the WoodScape data set is lower than that of the self-made data set, it still achieves good performance.

Table 4: Performance comparison of each algorithm on the WoodScape dataset.

| Model | mAP@0.5 | Precision | Recall |
|---|---|---|---|
| YOLOv5 | 85.20 | 83.10 | 76.60 |
| Ours | 86.03 | 83.51 | 77.56 |

4) Part of The Detection Effect Comparison Experiment

An example of the detection effect part on the data set built in this paper and WoodScape data set is shown in Figure 8. As shown in (a) and (c), in the underground garage with poor light, the proposed algorithm reduces the missed detection rate compared with the original YOLOv5 algorithm. Figure (d) At the edge of the fisheye image with large distortion, the original YOLOv5 algorithm cannot detect the distorted vehicle, while the algorithm is capable of accurately detecting distorted vehicles. The false detection of the original YOLOv5 in Figure (b). Therefore, this algorithm demonstrates good accuracy and robustness across various scenarios.

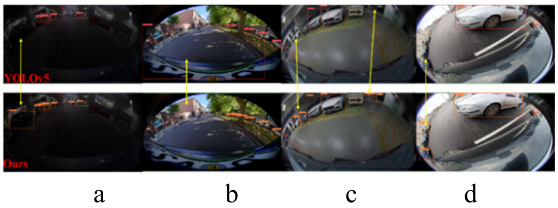

Figure 8: Example of comparison of detection effects.

## 5 CONCLUSIONS

In this paper, a series of improvement measures are proposed to facilitate the adaptive learning of distorted information DCNC3 module, add coordinate attention mechanism, and design Slim-Neck reconstruction feature fusion network from the aspects of feature learning, feature fusion, sample weight allocation and information transmission mode. The experimental outcomes demonstrate that the algorithm improves all indicators on both the dataset constructed in this paper and the public dataset. It not only effectively boosts the detection precision of vehicles in fisheye images, but also reduces the missed detection and false detection rate. However, the data set constructed in this paper is based on the urban parking environment, and the data set samples are not rich enough. It will be supplemented in the future.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

## ACKNOWLEDGMENTS

This work was financially supported by Natural Science Foundation of Fujian Province fund (Grant No. 2022J011247).

## REFERENCES

Lee M, Kim H, Paik J. Correction of barrel distortion in fisheye lens images using image-based estimation of distortion parameters[J], *IEEE ACCESS*, 2019, 7:45723-45733.

Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C], *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: 580-587. https://doi.org/10.1109/cvpr.2014.81

Yang W, Li Z, Wang C, et al.A multi-task Faster R-CNN method for 3D vehicle detection based on a single image[J], *Applied Soft Computing*, 2020, 95:106533. https://doi.org/10.1016/j.asoc.2020.106533

Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. *Computer Vision and Pattern Recognition*, 2018, 87(8): 101-104. https://doi.org/10.48550/arXiv.1804.02767

Chen X,Yu J,Wu Z.Temporally Identity-Aware SSD With Attentional LSTM[J], *IEEE transactions on cybernetics*, 2019,50( 06) : 2674-2686.

Jun Jiang, Donghai Zhai. Single-stage object detection Algorithm based on dilated convolution and Feature enhancement [J], *Computer Engineering*, 2021, 47(7):232-238+248.

Wei X, Wei Y, Lu X. RMDC: Rotation-mask deformable convolution for object detection in top-view fisheye camera[J], *Neurocomputing*, 2022, 504: 99-108. https://doi.org/10.1016/j.neucom.2022.06.116

Fremont V, Bui M T, Boukerroui D, et al. Vision-based people detection system for heavy machine applications[J], *Sensors*, 2016, 16(1): 128. https://doi.org/10.3390/s16010128

Yogamani S, Hughes C, Horgan J, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving[C], *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019: 9308-9318.

Hou Q , Zhou D , Feng J . Coordinate attention for efficient mobile network design[C], *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021: 13713-13722. https://doi.org/10.1109/cvpr46437.2021.01350

Chollet F. Xception: Deep learning with depthwise separable convolutions[C], *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 1251-1258. https://doi.org/10.1109/cvpr.2017.195

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS