Intelligent Perception System of Moving Robot for Dynamic Environment

Mingjin Weng*

College of Engineering Physics, Shenzhen Technology University, Shenzhen, China

*Corresponding author: 202300803104@stumail.sztu.edu.cn

Abstract. Object detection poses a significant challenge in the realm of computer vision for mobile robot grasping tasks. The rapid advancements in deep learning have led to the emergence of numerous novel object detection algorithms, resulting in improved accuracy in large-scale environments. Mobile robot grasping tasks demand highly responsive target detection, as any delay could hinder the robot's ability to react appropriately, thus impacting task execution. Additionally, the constantly changing background environment caused by a fast-moving robot introduces dynamic interference. To address these issues of real-time processing and dynamic background interference in mobile robot grasping tasks, it is crucial to enhance target detection speed, comprehensively capture features, and interpret image context information. This paper introduces a new information fusion structure based on the Yolov8 model and Gold-yolo aggregation distribution mechanism. The sampling approach retains the traditional Faster Implementation of CSP Bottleneck with 2 convolutions module (C2f) from YOLOv8 while incorporating the variable convolution module from Data Communication Network (DCN2). These two sampling methods are interwoven, enhancing the accuracy of detecting various shape information of the same object as the robot moves.

Keywords: Deep learning; Gold-yolo; DCN2.

1. Introduction

Traditional Conventional object detection relies on traditional computer vision algorithms and manually designed feature extractors. However, the advent of deep learning has led to the emergence of numerous novel object detection algorithms. These deep learning-based approaches offer superior accuracy in large-scale environments and enable end-to-end training, surpassing traditional methods. By integrating deep learning models and optimizing algorithm functions, the feature extraction process in detection tasks can be significantly enhanced, particularly for varying environments or dynamic targets.

The Region-CNN (R-CNN) series model, pioneered by Ross Girshick, achieves high-precision object detection through candidate region generation and convolutional neural networks [1]. However, it struggles with small object detection, resulting in low accuracy and slow detection speed. Joseph Redmon developed the YOLO series, which excels at quickly recognizing small objects [2]. Wei Liu's Single Shot MultiBox Detector (SSD) model combines the strengths of R-CNN and YOLO, markedly improving small object detection performance [3]. The Google Brain team introduced the transformer, featuring a multi-head attention mechanism [4]. Building on TopFormer theory, Huawei Group proposed a new aggregation-distribution (GD) mechanism for efficient information exchange in YOLO[5,6].

In mobile robot grasping tasks, real-time target detection is crucial. Any delay can hinder the robot's ability to respond appropriately, potentially compromising task execution. Additionally, a rapidly moving robot causes constant changes in the background environment, leading to dynamic background interference. To address these challenges of real-time processing and dynamic background interference, it is essential to enhance target detection speed, comprehensively capture features, and understand image context information.

This research employs a novel information fusion structure and feature extraction module to bolster image context information and thoroughly capture image features. The study utilizes a combination of the YOLOv8 model and Gold-yolo model to further improve the cross-layer fusion

capability of target detection information. Additionally, it incorporates C2f and DCN2 modules in the feature extraction process for information sampling.

2. Organization of the Text

2.1. Optimization of YOLOv8-gold object detection model

2.1.1YOLOv8 model

The YOLOv8 model is set to become open source in 2023, building upon its predecessors in the YOLO series [7]. It incorporates an enhanced C2F structure with gradient flow, which enhances model efficiency by reducing backbone computational requirements while boosting overall performance. The model's Neck component utilizes FPN and PAN techniques to consolidate information from various scales, enabling improved handling of multi-scale targets. In the Head section, a dual-pathway architecture separates classification and detection tasks, leading to enhanced detection accuracy.

2.1.2 Gold-yolo model

The traditional YOLO series model adopts the CSP mode in the neck part (the input feature map is divided into two branches, one branch is processed directly, the other branch is processed through a series of convolution and pooling operations, and the results of the two branches are merged at last) for information transmission [8,9]. However, information loss cannot be avoided in cross-layer transmission. Gold-yolo proposed a new aggregation-distribution (GD) mechanism to solve the problem of information loss in FPN mode. The feature alignment module (FAM) was used to collect feature maps of different sizes in backbone and align them by up-sampling or down-sampling. The feature information fusion module (IFM) is used to generate global features for the aligned features. Finally, the feature information distribution module (Inject) is used to distribute global features to all levels. Different tests have shown that Gold-YOLO shows higher detection accuracy and robustness when dealing with small targets and complex scenes

2.1.3 Information fusion structure

The Gold-yolo model replaces PANet's upsampling fusion stage with Low-GD and its downsampling fusion stage with High-GD in the YOLO series. This model introduces an information fusion structure that integrates the aggregation-distribution (GD) mechanism from Gold-yolo with the CSP structure from traditional YOLO, while maintaining CF2 for feature map downsampling. Unlike the conventional CSP structure, which splits the input feature map into two branches - one directly processed and the other reprocessed through convolution and pooling before merging - this new structure fuses multi-layer feature maps with specifically processed feature maps to create a new feature map. It then incorporates feature maps from other specific processes for multiple fusions. As illustrated in Fig. 1, feature maps from layers 2, 4, 6, and 9 are combined using simfusiom and IFM modules to form feature map Z. This Z map is subsequently injected during the information fusion of layers 4, 6, and 9, facilitating information exchange between interval layers and minimizing information exchange-related losses.

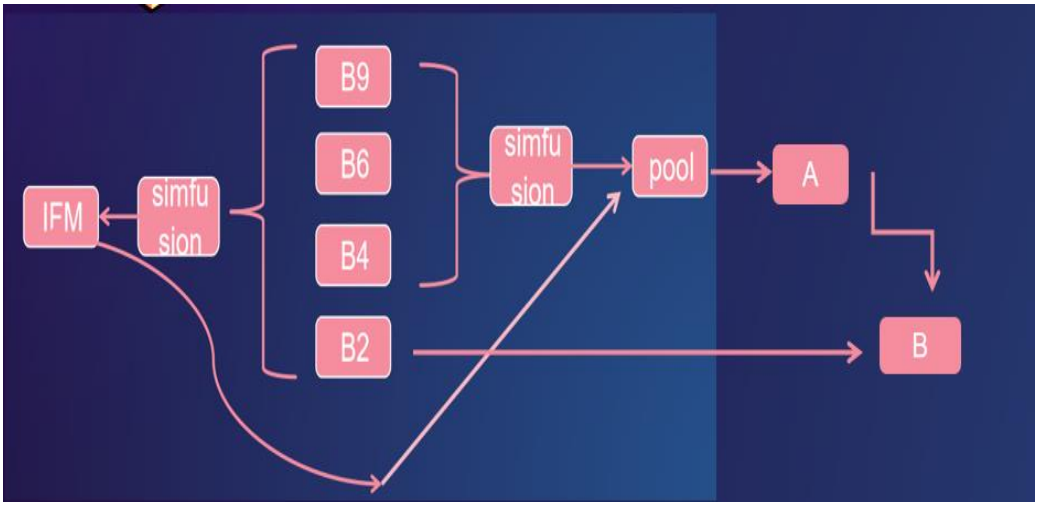

Fig.1 Model Structure Diagram

2.1.4 DCN2 module

In the traditional convolutional neural network, the fixed-shape convolution kernel is difficult to recognize different forms of the same object. Therefore, the deformable convolution kernel is used in the DCN2 module, so that the convolution kernel has the optimal structure in different stages of different feature maps. By learning a modulation scalar to control the offset range of the sampling points, the new sampling points are more concentrated in the region of interest. In the sampling part of traditional YOLOv8, the classical cf2 module and DCN2 module are used for interleaved adoption, which enhances the ability of the convolutional network to capture the information of the feature map.

2.2. Experiment

2.2.1 Dataset

Unmanned aerial vehicle (UAV) is a classic representative of mobile robots. This paper uses the UAV data set VisDrone to test the performance of model training. VisDrone data set is a large UAV perspective data set open source by Tianjin University and other teams. Drones equipped with cameras (or general drones) have been rapidly put into a wide range of applications, covering many fields such as agriculture, aerial photography, rapid delivery and surveillance. The dataset provides the following 12 classes: "Ignore area", "pedestrian ', 'people', 'bicycle, car, van, truck, tricycle, awning - tricycle,' bus', 'motor', 'others', where the ignored region and others are not valid target regions. In this paper, the data is designed as training set, validation set and test machine according to the ratio of 6:2:2.

2.2.2 Experimental parameters and evaluation metrics

The experimental setup utilizes Pytorch1.11.0, Python3.8, Cuda11.3, and an RTX3080Ti graphics card. For the FloW-Img dataset, 300 training iterations are employed with a batch size of 8 and image dimensions of 640×640. The VisDrone dataset uses 150 iterations and a batch size of 4, while the AUAIR dataset employs 200 iterations with a batch size of 8. The WiderPerson dataset undergoes 200 training iterations with a batch size of 4. All other parameters remain at their default values.

To evaluate the model's performance objectively, precision (P), recall (R), and mean average precision (mAP) are utilized as key metrics. Precision measures the ratio of correctly identified positive samples to the total detected samples, indicating the model's accuracy. Recall represents the proportion of actual positive samples successfully recognized by the model, reflecting its ability to identify relevant instances. Mean Average Precision (mAP) serves as a comprehensive evaluation metric, with higher values indicating superior model performance. In mAP50:5:95, "50" denotes the mAP calculated using an IoU threshold of 0.5, while "5:95" represents the average mAP values computed with IoU thresholds ranging from 0.5 to 0.95, incrementing by 0.05.

2.2.3 Ablation experiment

In order to verify the effectiveness of the information fusion structure and the c2f and DCN2 combining module, the following four sets of ablation experiments were performed. In the table, "√" means using this module, "--" means not using this module, and the results with the best performance are shown in bold boldface. The detailed data are shown in Table 1. Method 1 directly uses the YOLOv8 benchmark model to conduct experiments. In method 2, the accuracy, recall, mAP50 and MAP50:5:95 of the new feature extraction module combined with c2f and DCN2 modules are not changed on the YOLOv8 benchmark model. Method 3 introduces a new information fusion structure on the YOLOv8 benchmark model, and compared with method 1, mAP@50:5:95 is increased by 0.1%, and the recall and precision are basically the same as mAP50. Method 4 adds a new feature extraction module combined with c2f and DCN2 modules and a new information fusion structure to the YOLOv8 benchmark model, and the accuracy is significantly improved by 3.7%, and the recall rate is improved by 1.4%. mAP50 and mAP@50:5:95 also saw small improvements of 1.9% and 1.6%, respectively. Experiments show that adding the fusion structure or feature extraction module alone cannot effectively improve the performance of the model. When the fusion structure or feature extraction module is added at the same time, all indicators in the training set are significantly improved.

Table 1 Module ablation results

YOLOv8	Gold YOLO	C2f	DCN	Precision	recall	mAP50	mAP5:5:95
√	--	--	--	0.436	0.033	0.331	0.193
√	--	√	√	0.436	0.033	0.331	0.193
√	√	--	--	0.436	0.033	0.331	0.194
√	√	√	√	0.473	0.344	0.350	0.21

3. Conclusion

In the mobile robot grasping task, a convolutional neural network to achieve target detection is a mature vision scheme. In order to solve the detection problem of mobile robots, it is necessary to improve the speed of object detection, and strengthen the ability to comprehensively capture features and understand image context information. Based on the Yolov8 model, this paper proposes a new information fusion structure by combining the aggregation and distribution mechanism of Gold-yolo. In terms of sampling, the traditional c2f module of YOLOv8 is retained, and the variability convolution module of DCN2 is added. The two perform interleaved sampling, which improves the accuracy of reading different shape information of the same object in the process of robot movement. The new information fusion structure changes the traditional CSP structure of YOLOv8 under the premise of retaining the basic framework in YOLOv8, and adds multiple new information fusion structures to the information exchange of the feature map, which effectively strengthens the ability of information exchange between different layers in YOLOv8. The improved algorithm model has a significant improvement in the data set.

The combination of the new information fusion structure and the DCN2 module effectively improves the precision and recall of the model, especially when dealing with small objects and complex scenes. However, despite the remarkable progress, there are still some aspects that need to be further investigated and optimized. Firstly, although the new information fusion structure proposed in this paper performs well in the experiments, its high computational complexity may lead to insufficient real-time performance. Future research can consider to further optimize the computational efficiency under the premise of ensuring the performance to meet the real-time requirements of mobile robots. Secondly, the current model is mainly trained and tested for static or slowly changing background environments, and the robustness of the model still needs to be verified and improved for fast-moving or highly dynamic background environments. Therefore, future work

can explore more robust feature extraction methods and dynamic background modeling techniques to improve the adaptability of the model in complex environments.

References

[1] Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[2] Uijlings J R, van de Sande K E, Gevers T, Smeulders A W. Selective search for object recognition. International Journal of Computer Vision, 2013, 104: 154, doi: 10.1007/s11263-013-0634-5.

[3] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg A C. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-46448-0_2.

[4] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems, 2017, pp. 5998-6008.

[5] Wang Z, Zhang H, Li Y, Liu J. Gold YOLO: Towards Real-Time Object Detection with High Accuracy and Efficiency. International Journal of Computer Vision, 2023, 129(5): 1748-1764, doi: 10.1007/s11263-023-03079-y.

[6] Zhu J, Hu T, Zheng L, Zhou N, Ge H, Hong Z. YOLOv8-C2f-Faster-EMA: An Improved Underwater Trash Detection Model Based on YOLOv8. Sensors (Basel), 2024, 24(8): 2483, doi: 10.3390/s24082483.

[7] Tang X, Zhao S. The application prospects of robot pose estimation technology: exploring new directions based on YOLOv8-ApexNet. Frontiers in Neurorobotics, 2024, 18: 1374385, doi: 10.3389/fnbot.2024.1374385.

[8] Howard A G, Sandler M, Chen L C, Chen Z, Deng J, Rubner M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv preprint arXiv:1801.04381, 2018.

[9] Liu W, Wang Y, Song Q, Chen S. VisDrone 2019: A Benchmark and Dataset for Multi-Object Detection in Aerial Imagery. arXiv preprint arXiv:1910.01177, 2019.