Rotator-YOLOv5: Improved YOLOv5 for Vehicle and Vessel Detection in UAV Images

Yuxuan Zhang, Shuimiao Du
Sino-European School of Technology of Shanghai University (UTSEUS), 200444, China;
zyuxuan1316@gmail.com
dushuimiao@shu.edu.cn

Hengxiang He
Shanghai Aerospace Electronic Technology Institute, 201108, China.
hehengxiang0@163.com

Abstract*—In recent years, the widespread application of UAVs in national defence, military, and other fields, coupled with rapid technological advancements, has led to closer integration of UAVs and computer vision technology. Aerial images captured by UAVs hold significant potential for data mining; however, these images often feature arbitrarily rotated targets and complex backgrounds, resulting in insufficient detection accuracy. To address this challenge, we propose the Rotator-YOLOv5 algorithm, an enhancement of YOLOv5. (1) We integrate Circular Smooth Label (CSL) into YOLOv5 to achieve angle prediction with reduced computational complexity. (2) We design a two-branch structure that combines Convolutional Networks and Transformers, enabling feature information extraction from local and global perspectives. (3) By incorporating Partial Convolution (PConv) and RepNCSPELAN4, we further enhance the speed of our algorithm. Experimental results demonstrate that Rotator-YOLOv5 achieves a mean Average Precision (mAP) of 44.5% and a mAP_0.5 of 80.9% on UAV aerial images with an input size of 1024x1024, outperforming YOLOv5s by 1.1% and 6.1% respectively while maintaining real-time inference speed. Due to its lower deployment cost, Rotator-YOLOv5 is well-suited for real-time target detection tasks on embedded terminals with vertical viewpoints, such as robotic arms and UAVs.

Keywords—*Image processing, Object detection, Multilayer neural network, Supervised learning, Artificial intelligence*

I. INTRODUCTION

Drones are rapidly gaining prominence in various fields due to their affordability, ease of operation, and compact size. They can effectively monitor areas that are difficult for humans to access, such as volcanic craters and marine environments, while playing a crucial role in managing and surveilling high-risk sites, including construction zones and transportation networks. Additionally, drones are pivotal in terrain mapping, defense security, and military surveillance. With technological advancements, UAVs are increasingly integrated with smart technologies, enabling them to work with ground-based systems to automatically detect irregularities at construction sites and in urban traffic and perform autonomous detection and tracking tasks in military operations. Consequently, technological research in related fields holds vast application potential and market prospects.

Intelligent analysis of UAV images relies on computer vision technology, with target detection at its core, aiming to identify and locate objects in images accurately. The precision of target detection directly influences the overall system's intelligent performance. Despite significant advancements in target detection technology for natural images, the substantial differences between natural and UAV images present challenges. Specifically, targets in UAV-acquired images often exhibit varying rotation angles and sizes from different viewpoints, which conventional target detection algorithms frequently fail to accommodate, resulting in suboptimal recognition outcomes. However, numerous research efforts have been made to address target detection in UAV images, and issues such as inefficient training, low accuracy, and poor real-time performance persist, making this a challenging research area.

With the rapid development of UAV technology, the quality and resolution of images have improved significantly, and the number of detectable targets has increased dramatically. The typically high shooting altitude results in smaller target sizes in images, posing greater demands on the detector's performance in small target recognition scenarios. This paper focuses on developing a first-class fast detector for UAV image target detection, aiming to enhance the performance of UAV image target detection and advance its application in industrial settings. The contributions of this paper can be summarized as follows:

1. We propose a single-stage target detection algorithm to improve UAV image target detection accuracy and speed.
2. We integrate the Circular Smooth Label (CSL) method with YOLOv5, which reduces the number of parameters and computational load while achieving more accurate angle prediction than traditional regression methods.
3. We designed a two-branch structure combining Convolutional Networks and Transformers, replacing the Bottleneck module in YOLOv5's C3 component to maximize the use of both local and global feature information from the images.
4. We introduce Partial Convolution (PConv) and incorporate the RepNCSPELAN4 module from YOLOv9 to reduce the model's computational demands without decreasing accuracy.

II. RELATED WORK

Classical target detection typically employs Horizontal Bounding Boxes (HBB) to identify the location of targets within images. The most effective target detectors are often based on datasets labeled with horizontal boxes, such as ImageNet and MS COCO. Key metrics in target detection are detection accuracy and speed, which drive ongoing research and development in this field.

Deep learning-based target detection algorithms can be categorized into two-stage and single-stage detectors. Two-stage detectors, exemplified by R-CNN[1-2], have evolved into more advanced versions, such as Fast R-CNN[3] and Faster R-CNN[4]. These detectors perform candidate region extraction during detection, resulting in higher accuracy at

the expense of increased memory usage and slower detection speeds[5].

In contrast, single-stage detectors like YOLO[6], SSD[7], and RetinaNet[8] are renowned for their speed and simplicity. These detectors assess the location and category of targets directly during the forward propagation of a single network without the need for pre-generating candidate regions. This approach is advantageous for real-time applications and scenarios with limited computational resources, such as video content detection, mobile devices, and edge computing.

Several factors influence detection speed, including network structure, the number of parameters in the network backbone, and computational complexity. Networks like VGG, which demonstrated excellent performance in the ImageNet Image Recognition Challenge, showed that increasing network depth can significantly enhance image recognition accuracy. ResNet is another widely used backbone network known for its performance and flexibility, with various versions like ResNet50[9] incorporating a bottleneck structure to increase depth while minimizing complexity. Lightweight networks such as EfficientNet and CSPDarknet53 improve feature extraction capabilities by optimizing network structure. Methods like MobileNet[10] and ShuffleNet[11] reduce backbone complexity through improved computational and propagation strategies.

Feature fusion techniques, such as top-down Feature Pyramid Networks (FPN)[12] and bottom-up Path Aggregation Networks (PAN)[13], have also been proposed to enhance detector accuracy with minimal computational overhead. As a leading example of single-stage detectors, the YOLO series excels in speed and efficiency, making it ideal for real-time target detection. YOLO's core principle is to frame target detection as a regression problem, directly predicting bounding box coordinates and category probabilities from the image, thus enhancing detection speed while maintaining high accuracy.

Most detectors that output a rotating prediction frame include angle prediction in the model output and add a corresponding angle loss term to the loss function. Several rotating target detectors have improved classical detectors for more accurate angle prediction in recent years. For instance, RRPN[14] is a text detection framework based on the Faster R-CNN architecture that supports arbitrary rotation angles. It employs a Regional Proposal Network (RPN) to generate initial text proposals with angle information. Although this design increases computational effort, it simplifies angle regression.

The RoI-Transformer[15] is a model based on the Transformer architecture, specialized for rotating target detection. It enhances accuracy for rotating targets by introducing a Rotated Region of Interest (RRoI) pooling layer and utilizing rotated region of interest alignment. Oriented R-CNN[16] improves the design of RRoI, providing more reliable detection results in complex rotating scenes. SCRDet[17] focuses on small target detection, proposes a specialized loss function, and incorporates an attention mechanism to enhance the feature information in RRoI relevant to target recognition. R3Det[18] introduces two attention mechanisms, channel and spatial attention, which help the model focus on key features more efficiently. Rotated-RetinaNet further optimizes Focal Loss to reduce the weight of easy-to-predict samples and enhance the learning of difficult samples.

The periodic nature of the rotation angle in images can lead to boundary problems in regression tasks, causing unstable convergence and slow model training. The models above rely on regression for angle prediction and do not address boundary issues. DarkNet-RI[19] uses a pixel-by-pixel classification method to predict the location of rotating targets by classifying each pixel in the image, which avoids the regression problems of traditional methods. However, pixel-by-pixel classification increases computational complexity, especially when handling high-resolution images. Circular Smooth Label (CSL)[20] addresses the angle prediction as a classification problem and combines it with a window function to resolve periodicity issues caused by the rotated frame. This method is easy to integrate and does not significantly increase model parameters.

III. METHOD

To improve the detection of rotating targets in UAV images, we developed Rotator-YOLOv5 based on the YOLOv5 framework. In this section, we first present an overview of the Rotator-YOLOv5 architecture. We then provide a detailed explanation of its key components, including the Partial Convolution (PConv) module[21], the RepNCSPELAN4 module[22], and the CTR3_Mix module.

A. *Overview of Rotator-YOLOv5*

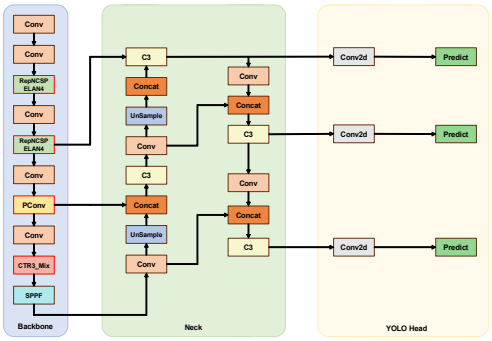


Fig. 1 The architecture of Rotator-YOLOv5

Fig 1 illustrates the structure of our proposed Rotator-YOLOv5 network. Compared to YOLOv5, Rotator-YOLOv5 incorporates five key modifications. Specifically, we replace the C3 module in the Backbone with various alternative modules. Additionally, we have adjusted the Head layer Predict module size from [256, 512, 1024] to [256, 256, 512] to better accommodate the detection of numerous small targets present in UAV images.

B. *PConv Module*

In convolutional neural networks, there is often a significant amount of highly similar image information across different channels. Many feature map channels may contain redundant information, leading to unnecessary repetitive processing during forward propagation without contributing additional useful content. The Partial

Convolution (PConv) module, proposed by FasterNet, addresses this issue by reducing computational redundancy and memory accesses. Its core principle is to perform convolution operations on only a subset of the input channels, thereby decreasing computational load while still effectively extracting spatial feature information.

To enhance the network's efficiency in utilizing feature information, we incorporated the PConv module into the Backbone layer of the Rotator-YOLOv5 network, replacing the C3 module in the 6th layer. The PConv module selectively applies regular convolution operations to certain input channels, leaving other channels unprocessed. For sequential or regular memory accesses, it computes only the first or last continuous subset of channels, assuming these are representative of the entire feature map.[23] This approach maintains the same number of channels in the input and output feature maps, ensuring no loss of generality while significantly improving computational efficiency. This results in:

The FLOPs of PConv: $h \times w \times k^2 \times c_p^2$

where h and w denote the width and height of the feature map, k is the size of the convolution kernel, and c_p is the number of channels on which the regular convolution acts, and in the usual case, $\frac{c_p}{c} = \frac{1}{4}$, so that PConv produces only $\frac{1}{16}$ as many FLOPs as the regular convolution.

Number of memory accesses for PConv: $h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p$

We can learn that the memory access of PConv is $\frac{1}{4}$ of that of regular convolution.

In FasterNet Block, the PConv layer is usually followed by a Conv(1x1) layer, and our network follows this structure

C. RepNCSPELAN4 Module

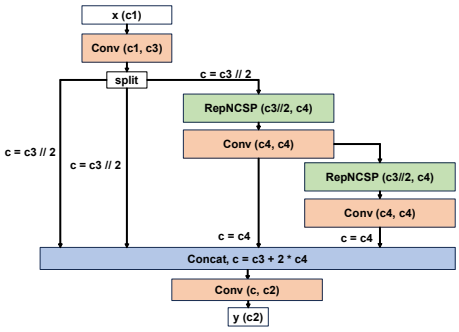


Fig. 2 The architecture of the RepNCSPELAN4 Module

RepNCSPELAN4 is a feature extraction and fusion module used in YOLOv9, functioning similarly to the C3 module in YOLOv5. This module integrates the CSPNet and ELAN architectures for gradient path planning, aiming to create a Generalized Efficient Layer Aggregation Network (GELAN) that balances lightweight design, inference speed, and accuracy. By replacing the original ELAN with GELAN incorporating CSPNet blocks and utilizing RepConv as the computational unit, RepNCSPELAN4 is restructured into RepN-CSP-ELAN4. The module's detailed structure is illustrated in Fig 2.

D. CTR3_Mix Module

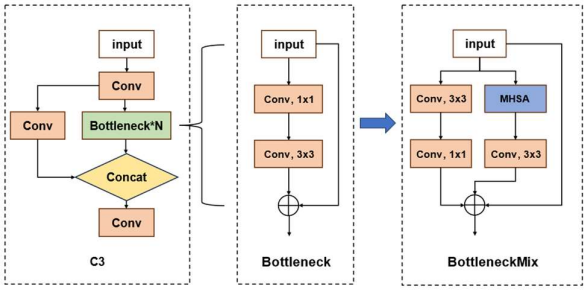


Fig. 3 The architecture of C3, Bottlenck and BottleneckMix

As the depth of convolutional neural networks increases, training difficulties arise, and simply adding more convolutional layers may not necessarily enhance detection performance.[24] The Residual Block, proposed in ResNet, addresses this by introducing a "constant mapping" that allows the input to flow directly to the output layer. This mechanism enables efficient gradient transfer even in deep networks, preventing gradient vanishing issues. By incorporating residual blocks, neural networks become easier to optimize and achieve higher accuracy with significantly increased depth.

BotNet enhances ResNet by integrating Transformers, replacing some convolutional layers with Multi-Head Self-Attention (MHSA). This modification improves network performance with minimal additional computational effort. While convolutional layers focus on perceiving target features in small areas, Transformers can capture features over larger areas. For UAV target detection, it is crucial to observe both the overhead view features of the target and the complex background information.[25]

Inspired by this, we improved the BottleNeck layer in the C3 module of YOLOv5 and proposed the BottleNeckMix module. This new module combines convolutional layers with MHSA and adopts a dual-branch structure instead of a single convolutional layer. The structure is illustrated in Fig 3.

IV. Experiment

This section introduces the dataset used for the experiments, the "Shen Ji Miao Suan" dataset released by the Shenyang Institute of Aeronautics and Astronautics (SIAS). We then describe the evaluation metrics and implementation details of the experiments. The experiments focus on comparing the performance of Rotator-YOLOv5 with other algorithms, providing a comprehensive analysis of its effectiveness.

A. 4.1 Dataset

In this paper, we utilize the UAV image dataset released by the Shenyang Institute of Aerospace Industry (SIAS) for the 2023 "Shen Ji Miao Suan" algorithm competition to evaluate the performance of our algorithms. The dataset comprises 10,000 images and 135,927 instances of UAVs captured from an overhead view, including 6,338 images taken under natural light and 3,662 infrared images. The

images have resolutions of 1920 x 1080 pixels for visible light and 640 x 512 pixels for infrared. The primary targets for recognition in the dataset are cars and boats, with 109,576 instances of cars and 26,351 instances of boats. The dataset is divided into a training set with 8,000 images and 95,810 instances and a validation set with 2,000 images and 40,117 instances. This division allows for a robust assessment of the algorithm's performance on both seen and unseen data.

B. 4.2 Evaluation Metrics

We adopt commonly used evaluation metrics, namely P(precision), R(recall), mAP (mean average precision), and $mAP_{0.5}$ (mean average precision at IOU = 0.5) in the experiments. P and R are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

where TP represents the number of positive samples correctly identified, FP represents the number of negative samples considered positive, and FN represents the number of positive samples considered negative. AP (average precision), mAP, and $mAP_{0.5}$ need to be computed based on P and R.

$$AP = \int_0^1 P(R)dR \tag{3}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i\,, IOU = 0.5 : 0.05 : 0.95 \tag{4}$$

$$mAP_{0.5} = \frac{1}{N}\sum_{i=1}^{N} AP_i\,, IOU = 0.5 \tag{5}$$

N denotes the number of categories of the target.

C. 4.3 Implementation Details

The implementation of Rotator-YOLOv5 utilizes PyTorch (version 1.10.1) as the underlying framework and runs on the Ubuntu 20.04 operating system. An NVIDIA RTX 3060 Ti GPU with 8GB of RAM is employed for training and testing. During training, an SGD optimizer with momentum set to 0.98 and weight decay set to 0.001 is used. A warmup strategy is implemented to enhance training stability, where the learning rate gradually decreases to 0.001 over the first five epochs and then continues at this rate. Due to hardware limitations, the image size is also adjusted to 1024x1024 pixels, and the batch size is set to 8.

Other models, including YOLOv5, YOLOv7, Oriented R-CNN, Rotated RetinaNet, and ROI Transformer, were trained and tested under the same conditions as Rotator-YOLOv5. All models used an image size of 1024x1024, and the default parameter settings from the respective reference articles were applied.

D. 4.4 Experimental results

We evaluate Rotator-YOLOv5 in terms of P、R、$mAP_{0.5}$、mAP, and model computation $FLOPs$.To demonstrate the strengths of our proposed algorithm, we compare it with YOLOv5s, YOLOv7, YOLOv7-tiny, Oriented R-CNN, Rotated RetinaNet, Roi Transformer. The results of the experiments on the " Shen Ji Miao Suan " dataset are shown in the following table:

TABLE 1. COMPARISON RESULTS OF ROTATOR-YOLOv5 AND OTHER ALGORITHMS ON THE DATASET

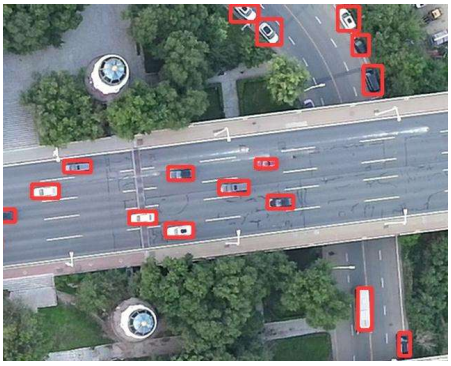
Method	?? (%)	?? (%)	$?????_{0.5}$ (%)	????? (%)	FLOPs (G)
YOLOv5s	79.4	69.0	74.8	43.4	15.8
YOLOv7	74.3	65.4	69.4	38.1	103.2
YOLOv7-tiny	77.0	67.1	71.6	38.0	**13.0**
Oriented R-CNN	77.8	41.6	65.0	32.5	211.4
Rotation Faster R-CNN	72.7	42.4	63.6	28.4	211.3
Rotation RetinaNet	76.2	42.2	60.8	25.8	210.0
Roi Transformer	78.3	43.2	68.3	33.0	225.2
Rotator-YOLOv5	**81.4**	**73.9**	**80.9**	**44.5**	14.9

Table 1 shows the performance of the eight target detection algorithms through different metrics, Rotator-YOLOv5 has a more balanced performance in terms of accuracy and computation, with 80.9% and 44.5% for our methods mAP_0.5 and mAP respectively. It is higher than any other method, and in terms of computation volume, it is only 1.9G more than YOLOv7-tiny with extremely good accuracy performance.

TABLE 2. COMPARISON RESULTS OF ROTATOR-YOLOv5 AND OTHER ALGORITHMS ON TWO CATEGORIES OF “SHEN JI MIAO SUAN” DATASET

Class	YOLOv5s	YOLOv7	YOLOv7-tiny	Oriented R-CNN	Rotation Faster R-CNN	Rotation RetinaNet	Roi Transformer	Rotator-YOLOv5
Car	86.0	82.6	83.9	79.5	77.7	71.4	80.0	**91.3**
Ship	63.7	56.1	59.2	50.4	49.5	50.2	56.6	**70.5**

Table 2 demonstrates the average accuracies (mAP50) of the eight methods in both dataset categories. Overall, Rotator-YOLOv5 outperforms the other algorithms on car and boat classification tasks. In the target detection task for vehicles, mAP50 outperforms the second place (YOLOv5s) by 5.3% and the heavy detection task for ships by 6.8%. The comparison of the detection results of YOLOv5s and Rotator-YOLOv5 on the same image is shown in Fig. 4. Rotator-YOLOv5 was able to obtain the location and category information of the target to be detected by the image information of the exposed part even when the target (the vehicle under the viaduct) appeared to have a wide range of occlusion. In contrast, YOLOv5s did not detect this target. This shows that Rotator-YOLOv5 can detect the target with better discrimination in detail.



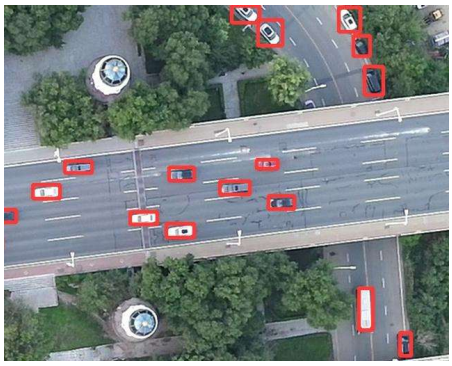


Fig. 4 The detection results on one image of the test set of the "Shen Ji Miao Suan" dataset:
(a) detection result of YOLOv5s **(b)** detection result of Rotator-YOLOv5

V. CONCLUSION

In this paper, we propose Rotator-YOLOv5, a rotating frame target detection algorithm that enhances the existing YOLOv5 by incorporating UAV image target detection characteristics and integrating it with Circular Smooth Label (CSL). Our approach combines Convolutional Neural Networks (CNN) and Transformers, leading to the development of a new YOLOv5 Bottleneck module, CTR3_Mix. This module leverages the local feature extraction capability of CNNs and the global perception capability of Transformers, effectively addressing the presence of small targets and complex backgrounds in UAV images, thereby significantly improving detection accuracy.

We also introduce the RepNCSPELAN4 module from YOLOv9 during the feature extraction stage. This module enhances the network's utilization of image features by exploiting multiple connections between convolutional layers, reducing computational effort without sacrificing detection accuracy. Additionally, we replace some of the C3 modules with Partial Convolution (PConv), which further decreases the model's computational load by eliminating redundant computations within the feature map.

Our proposed algorithm demonstrates superior accuracy and computational performance compared to existing detectors for UAV aerial images. Rotator-YOLOv5 requires less computational effort for large-scale image inputs than YOLOv5s, making it suitable for deployment on lightweight terminals alongside high-resolution imaging devices.

REFERENCES

[1] X.Xingxing, C. Gong, J. Wang, X. Yao, J. Han, "Oriented R-CNN for Object Detection" 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[2] R. Girshick, J. Donahue, T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation" Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[3] R. Girshick. "Fast r-cnn" Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[4] S. Ren, K. He, R. Girshick, "Faster r-cnn: Towards real-time object detection with region proposal networks". Advances in neural information processing systems, 2015, 28.

[5] Z. Zou, K. Chen, Z. Shi et al. "Object detection in 20 years: A survey" Proceedings of the IEEE, 2023, 111(3): 257-276.

[6] J. Redmon, S. Divvala, R. Girshick, "You only look once: Unified, real-time object detection" Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[7] Liu W, Anguelov D, Erhan D, et al. Ssd: "Single shot multibox detector" Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[8] T Y. Lin, P. Goyal, R. Girshick, "Focal loss for dense object detection" Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[9] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition"2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] G.A Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand Tobias, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". ArXiv, 2017, abs/1704.04861.

[11] X. Zhang, X. Zhou, M. Lin, "Shufflenet: An extremely efficient convolutional neural network for mobile devices" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[12] T Y. Lin, P. Dollár, R. Girshick, et al. "Feature pyramid networks for object detection" Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[13] W.Wang, E. Xie, X. Song, et al. "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network" Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8440-8449.

[14] J. Ma, W. Shao, H. Ye, "Arbitrary-oriented scene text detection via rotation proposals", IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[15] J. Ding, N. Xue, Y. Long, "Learning RoI transformer for oriented object detection in aerial images" Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2849-2858.

[16] X. Xie, G. Cheng, J. Wang, "Oriented R-CNN for object detection" Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3520-3529.

[17] X. Yang, J. Yang, J. Yan, "2019": 8232-8241.

[18] X.Yang, J. Yan, Z.Feng, "R3det: Refined single-stage detector with feature refinement for rotating object" Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 3163-3171.

[19] Z. Mohsen, E. Ali, M. Greenspan, "Oriented Bounding Boxes for Small and Freely Rotated Objects" IEEE Transactions on Geoscience and Remote Sensing, 2022.

[20] Y, Xue, Y, Junchi. “On the Arbitrary-Oriented Object Detection: Classification-based Approaches Revisited”. ArXiv, 2021, abs/2003.05597v3.

[21] J. Chen, S. Kao, H. He, “Run, Don't walk: Chasing higher FLOPS for faster neural networks”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12021-12031.

[22] C Y. Wang, I H. Yeh, H Y M. Liao, “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information”. arXiv preprint arXiv:2402.13616, 2024.

[23] H. He, Y. Qiao, X. Li, “Automatic weight measurement of pigs based on 3D images and regression network”. Computers and Electronics in Agriculture, 2021, 187: 106299