ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

Vehicle Detection in Aerial Drone Overhead Images based on YOLOv8

Shengkai Feng*, Xuejun Niu

School of people's public security of University, Beijing 100038, China

*Corresponding author: gadxnxj@sina.com

Abstract

Vehicle target detection technology is widely used in fields such as autonomous driving and intelligent traffic monitoring, but practical applications have strict requirements for its real-time detection, accuracy, and resource consumption. This article proposes the ACEW-YOLOv8 model for vehicle detection in drone overhead images. This model is based on YOLOv8n and replaces two convolutional layers with AKConv in the backbone network to enhance feature extraction capability; Add an ECA channel attention module to the Neck network to enhance feature expression and fusion capabilities; Adopting the Wise-IoUv3 loss function to reduce the impact of low-quality images on model training.

Keywords

YOLOv8; Drone Aerial Photography; Vehicle Inspection; AKConv; ECA; Wise - IoUv3.

1. Introduction

In recent years, in the fields of video detection, deep learning, and computer vision, vehicle object detection technology has become a hot research topic due to its broad application prospects, continuously promoting innovation and development of related technologies. In the research and development of auto drive system, vehicle target detection technology plays an indispensable role. Its accurate detection results provide a key basis for the decision-making and control of autonomous vehicle, which is directly related to the safety and reliability of driving. In the intelligent traffic monitoring system, vehicle target detection technology can monitor vehicles on the road in real time, achieve functions such as traffic flow statistics and violation recognition, effectively improve the intelligence level of traffic management, greatly improve operational efficiency, and significantly reduce labor costs.

However, in practical application scenarios, many tasks impose strict requirements on vehicle object detection technology. On the one hand, there is a high expectation for real-time and accurate detection to meet the needs of rapid response to changing road conditions during autonomous driving, as well as the precise recognition of vehicles in complex scenarios through intelligent traffic monitoring. On the other hand, the limitations of computing resources and energy consumption cannot be ignored, especially in some mobile devices or resource constrained environments. How to reduce computing and energy consumption while ensuring detection accuracy has become a key challenge that urgently needs to be overcome in this field.

With the continuous advancement of technology, the application of drones has become increasingly widespread. The use of drones for target detection plays an important role in areas such as emergency rescue, police investigation, agricultural development, and traffic detection. However, when using drones for object detection tasks, there are challenges such as mutual occlusion between targets, dense distribution of targets, small size, and low resolution [1]. In addition, drone images are usually small, and due to factors such as shooting distance, angle, and weather, the clarity of targets is low, which

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

reduces the recognition accuracy of drone target detection. Traditional algorithms are difficult to apply, so designing an efficient drone target detection model is particularly important.

The current object detection algorithms can be divided into two categories. One is the "two-stage" object detection algorithm based on candidate boxes, which first generates the target region and then classifies it through convolutional neural networks. Common two-stage detection algorithms include R-CNN [1], FastR-CNN [2], Faster R-CNN [3], etc; Another type is a one-stage object detection algorithm based on regression methods, which does not require the generation of pre selected boxes and directly extracts features in the network to predict the classification and position of objects. Common one-stage object detection algorithms include YOLO [4] series and SSD [5] series. The two-stage algorithm has a slow detection speed and is not suitable for real-time detection from the perspective of unmanned aerial vehicles;The single-stage algorithm has good real-time performance but low accuracy, especially for detecting small targets, which needs further optimization.

In the practical application environment of drone target detection, it is also necessary to consider the characteristics of model lightweight, detection accuracy, and easy deployment.In recent years, many researchers have done a lot of work in the field of unmanned aerial vehicle target detection, proposing many effective research methods. The YOLO series detection algorithms with good performance have also been applied to other target detection tasks.The UAV-YOLOv8 proposed by Wang et al. [6] introduces BiFormer into the backbone network and combines it with FasternBlock to focus on relevant attention through overlapping block embedding and feature transformation, without distracting the attention of other irrelevant labels. The FFNB module is designed to improve the accuracy of the model. In order to suppress information loss during long-distance feature transmission, Li et al. [7] replaced the C2f of YOLOv8 with GhostlockV2, which suppresses information loss during long-distance feature transmission while reducing the number of model parameters. In addition, WiseIoUloss was used as the bounding box regression loss to distinguish the quality of anchor boxes through outlier degree, focusing on general quality anchor boxes and not being affected by low-quality anchor boxes, thus improving the overall performance of the model. Zhai et al. [8] addressed the issue of fine-grained information loss caused by stride convolution and pooling layers. In the feature extraction stage, SPD Conv was used instead of Conv to extract multi-scale features. Spatial to depth layers and non stride convolution layers were employed to reduce the model's parameter count without sacrificing accuracy. Additionally, GAM attention mechanism was introduced in the neck section to enhance the model's fusion of target features. Cao et al. [9] introduced TripletAttention into Backbone to address the issues of low accuracy and slow speed in infrared small target detection. They utilized a three branch structure to capture cross dimensional interactions of data. In addition, the new model adopted the Slim Neck architecture and achieved a balance between accuracy and speed by introducing lightweight convolution GSConv. The DC-YOLOv8s model proposed by Lei et al. [10] adds a new small target layer to enhance the sensitivity of the model to small target scales, and adds an attention mechanism to the detection head to improve the localization performance of the model for small targets in various environments. Li Yanchao et al. [11] proposed CEM for Carafe upsampling operator to alleviate the impact of channel compression, and proposed SPM based on CA attention mechanism, which can more effectively learn the features of the interested position.

2. Improved Drone Aerial Images based on ACEW-YOLOv8 Model

2.1 AKCov Convolutional Kernel

2.1.1 Main Features of AKConv

YOLOv8, as a cutting-edge achievement in object detection technology, has demonstrated excellent performance in various complex environments and scenes.This algorithm can quickly adapt to the needs of different specific scenarios by fine-tuning pre trained models or applying transfer learning strategies.This feature not only ensures its efficiency, significantly shortens the development cycle, and reduces development costs, but also enables YOLOv8 to be quickly deployed in multiple fields, making it highly practical.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

The architecture of YOLOv8 consists of three core components: Backbone, Neck, and Head, each of which plays a unique and important role in object detection.

The input end is responsible for receiving and preprocessing image data, converting it into standardized sizes that the model can handle.This process ensures the consistency of input data and provides a stable foundation for subsequent model processing.The backbone network is constructed based on convolutional neural networks, and its main function is to conduct in-depth feature mining on preprocessed images.Through multi-layer convolution operations, the backbone network can extract rich feature information from images, providing strong support for subsequent target recognition and localization.The neck network plays an important role in integrating feature maps at different levels. By fusing features at different scales, it enhances the expressive power and propagation efficiency of features, further improving the detection accuracy of the model.Finally, the detection head performs target localization and category prediction tasks based on the feature information provided by the neck network, thereby generating the final detection result.

YOLOv8 provides five different scale model versions, labeled as n, s, m, l, and x, to meet the needs of different application scenarios and computing capabilities.Generally speaking, the larger the model size, the stronger its ability to process high-resolution images, extract richer feature information, and thus achieve higher detection accuracy.However, as the model size increases, the consumption of computing resources also increases accordingly, resulting in a decrease in running speed.Therefore, in practical applications, it is necessary to comprehensively consider various factors such as the size of the dataset, the actual situation of computing resources, and the specific requirements for processing data, and carefully select and adjust the model to achieve the best balance between model complexity and detection performance, ensuring that the model meets detection requirements while minimizing resource consumption.

2.1.2 Sub Heading

Firstly, define the initial sampling position: generate a regular sampling grid, then create irregular sampling grids for the remaining sampling points, and finally concatenate them together to form the overall sampling grid.The regular convolutional network (i.e. square size convolution kernel) takes the center as the collection origin, while the irregular collection network takes the top left corner as the collection origin. From this, the convolution operation at the corresponding point can be obtained:

$$Conv\ (P_0) = \sum_{P_n \in R} w_{p_n} \times x(P_0 + P_n) \tag{1}$$

R represents the generated sampling grid, x is the pixel of the corresponding point, and w is the corresponding convolution parameter.

Using the Linear Deformable Convolution (LDConv) method, for the irregular convolution kernel size generated by the above method, convolution operations are performed to obtain the corresponding offset, which is then added to the original coordinates to generate new sampling coordinates corresponding to the convolution.Convolution generates different sample shapes at different positions in the feature map.Obtain the corresponding position features through interpolation and resampling.Then apply the corresponding convolution operation to extract features.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{2}$$

Afterwards, the resampled features are transformed into a four-dimensional form (C, N, H, W), and then the features are extracted by applying a three-dimensional convolution (Conv3d) with a stride and convolution kernel size of (N, 1, 1).This process involves resampling features in the channel

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

dimension (C × N, H, W) using three types of 2D and 3D convolution kernels, followed by reducing the feature dimension to (C, H, W) through a 1x1 convolution operation.

2.2 CSPNeXt Structural Improvement

When conducting small object detection from the perspective of drones, existing models often face problems such as poor feature extraction capabilities and low efficiency in multi-scale information fusion.To effectively address these challenges, this paper proposes an improved YOLOv8n model based on the CSPNeXt backbone network mechanism.This improvement plan significantly enhances the model's adaptability to complex scenarios by optimizing the backbone network architecture and strengthening feature fusion capabilities.The specific improvement content is as follows:

The original backbone network of YOLOv8n is built based on C2f modules, which has the advantage of lightweight, but has certain limitations in global modeling of small target features.In view of this, this article adopts CSPNeXt to replace the original backbone network.CSPNeXt integrates the concept of Cross Stage Partial Network (CSPNet) and lightweight design, and improves feature extraction efficiency through optimization in the following aspects:

Use Split Concat structure to split the feature map into two parts.One part of it is directly passed on to the next stage, while the other part extracts local features through Depthwise Conv, and finally concatenates and fuses the two parts of the features.This design effectively alleviates the problem of gradient vanishing and significantly enhances the network's ability to capture detailed features.Introduce Ghost convolution in CSPNeXt module.Ghost convolution generates redundant feature maps through linear transformations, significantly reducing the number of parameters while ensuring feature representation capability.Experimental data shows that Ghost convolution can reduce the parameter count of a single module by 30%. Connect SPD Conv (Space to Depth Convolution) after each CSPNeXt module. By downsampling the spatial dimension and expanding it in the channel dimension, SPD Conv can effectively preserve fine-grained features and further enhance the fusion ability of multi-scale information.

The input image is first subjected to 3 × 3 standard convolution for preliminary feature extraction, and then sequentially downsampled through 4 CSPNeXt modules to finally output a high-dimensional feature map with a resolution of 32 × 32.The internal structure of CSPNeXt module includes core components such as Split branch, Ghost convolutional layer, and SPD Conv module.

2.3 ECA Channel Attention Module

2.3.1 Basic Characteristics of ECA

Drones typically perform missions at high altitudes, and the target objects in their captured images are often small in size and complex in shape. This places higher demands on the feature extraction and detail resolution capabilities of the model. In order to address the above issues, this paper introduces ECA (Efficient Channel Attention) channel attention mechanism in the front-end of YOLOv8's small object detection module to enhance the model's ability to extract key features.

The ECA module is a lightweight channel attention mechanism designed to dynamically adjust the weights of different channels in the feature map by modeling the dependencies between channels. Compared with the traditional Squeeze and Excitation (SE) module, the ECA module uses 1D convolution instead of fully connected layers, significantly reducing the number of parameters and computational complexity while maintaining the effectiveness of channel attention. By generating channel weights through nonlinear activation functions, the weights are ultimately multiplied with the original feature map to enhance the feature expression of important channels and suppress the features of unimportant channels, thereby improving the sensitivity of the model to target features.

2.3.2 Main Technical Methods of ECA

The core idea of ECA is to use a one-dimensional convolution to capture the dependency relationships between different channels, and there is no dimensionality reduction process in this process. As an improved version of the SE module, comparing the performance between different variants of the SE

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

module, it can be found that using a single fully connected layer in the SE module performs better, and avoiding channel reduction can greatly improve learning efficiency.

Firstly, input a C × H × W feature map, where C is the number of channels, H and W are the width and height, respectively. Perform global average pooling GAP on the given feature map to reduce the spatial dimension to 1 × 1, and obtain the global information of each channel.

In terms of the effect of cross channel interaction, considering the global GAP will make the model parameters too large and the complexity too high. In order to lightweight the model while considering module efficiency, each channel only interacts with k adjacent channels. The determination of k is an adaptive method, and the formula is as follows:

$$[k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}}] \tag{3}$$

Among them, γ and b are hyperparameters, C is the number of input channels, and odd represents an odd number that is rounded down and has an absolute value, ensuring that the number of convolution kernels is odd. The final output of this 1D convolution is the channel attention weight. This experiment uses 1 as the hyperparameter assignment.

Finally, the output of the 1D convolution is passed through the Sigmoid activation function to obtain normalized channel attention weights:

$$a_c = \sigma(\text{Conv1D}(z)_c) \tag{4}$$

Among them, σ is the Sigmoid function, $a_c$representing the attention weight of the c-th channel.

2.4 Improvement of W IOU Loss Function

The loss function of YOLOv8 algorithm mainly consists of category classification loss and bounding box regression loss, where the form of bounding box regression loss is DEF Loss+CIoU Los9s.

$$CIoU = IoU - (\frac{\rho^2(B^{pred}, B^{gt})}{c^2} + \alpha v) \tag{5}$$

$$v = \frac{4}{\pi}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^{pred}}{h^{pred}}) \tag{6}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{7}$$

In the formula, IoU represents the degree of overlap between the target box and the predicted box, β gt and β prd represent the diagonal lengths of the real bounding box and the predicted bounding box, respectively, and v represents the consistency of aspect ratio.

During the annotation process of the VisDrone2019 dataset, it is inevitable that there may be some target objects with poor annotation quality.If the model generates high-quality anchor boxes for these

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

objects, it will result in significant IoU loss.In order to reduce the adverse effects of low-quality anchor boxes on detection results, Wise-IoUv3 is used to replace the original loss function.Wise-IoUv3 utilizes a dynamic non monotonic focusing mechanism, replacing IoU with "outlier degree" for quality evaluation of anchor boxes, and also provides a reasonable gradient gain allocation strategy.This strategy reduces the competitiveness of high-quality anchor boxes while also minimizing the harmful gradient generated by low-quality examples.In this way, WIoU can focus on anchor boxes of ordinary quality, thereby improving the overall performance of the detector.

Wise-IoUv3 first calculates the intersection and union between the predicted box and the true box to obtain the basic IoU loss value, which is an indicator of the degree of overlap between two bounding boxes.It uses outlier degree to evaluate the quality of anchor boxes based on Wise-IoUv1, increasing the loss contribution of ordinary quality anchor boxes.Outlier degree is obtained by measuring the IoU loss value and the distance from the center point.Then calculate the dynamic non monotonic focusing coefficient based on the outlier degree to reduce the interference of low-quality anchor boxes on the overall training effect.In order to maintain a high level of gradient gain, Wise-IoUv3 introduces the mean IoU loss as a normalization factor and optimizes convergence speed by dynamically updating the normalization factor.

Outlier degree can reflect the quality of anchor boxes, and the smaller the outlier degree, the higher the quality of anchor boxes.The definition formula of outlier degree is as follows: When $\beta=\delta$ and the outlier degree of the anchor box satisfies $\beta=C$ (C is a constant value), the gradient gain of the anchor box reaches its maximum.

2.5 Overall Structure of ACEW-YOLOv8

Considering the practical application scenarios, this paper adopts YOLOv8n as the basic model for modification. AKconv is used to replace the two convolutional layers in the Backbone section, and ECA attention mechanism is added to the Neck layer to capture long-distance dependencies, improving the overall network's feature expression and fusion capabilities. Finally, the dynamic non monotonic loss function Wise-IoUv3 is used to reduce the impact of low-quality images on model training and enhance the performance of the model.

3. Experimental Results and Analysis

3.1 Introduction to Dataset and Experimental Platform

The VisDrone2019 dataset is a benchmark dataset for drone aerial photography developed by the AISKYEYE team at the Machine Learning and Data Mining Laboratory of Tianjin University, specifically designed for computer vision task research.This dataset contains 288 video clips (totaling 261908 frames) and 10209 static images, collected from different geographical regions of China, covering diverse scenes such as cities and rural areas, and including various weather conditions and lighting changes.All data is obtained through multiple types of drone cameras, ensuring the richness of environmental features and representativeness for practical applications.Its core value is reflected in the manually annotated 2.6 million fine target bounding boxes, covering typical target categories such as pedestrians, cars, bicycles, and tricycles, providing important support for object detection, single/multi object tracking, and crowd counting research from the perspective of drones.

The Unmanned Aerial Vehicle Detection and Tracking Dataset (UAVDT), published by ICCV, is a large-scale visual benchmark dataset for unmanned aerial vehicle platforms.This dataset contains 80000 annotated images, sourced from 10 hours of raw aerial video (resolution 1080 × 540, frame rate 30 FPS), covering six typical traffic scenarios including squares, main roads, and highways.Its annotation system is optimized for vehicle targets, with precise bounding box annotations included in each frame, along with attribute information such as vehicle category (such as cars, trucks) and occlusion degree.Through multi scene and highly dynamic visual data collection, UAVDT focuses on providing robust validation and performance improvement research for target detection and tracking algorithms in complex environments.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

The experimental platform is the Windows 11 operating system, the GPU is NVIDIA RTX3060 Laptop, the deep learning framework is Pytorch 1.13.1, the programming language is Python 3.9.13, and the CUDA version used is 11.6.

The parameter settings for this experiment are as follows: the image input size is 640 × 640, the batch processing size is set to 16, the initial learning rate is 0.01, the attenuation coefficient is set to 0.0005, the optimizer is SGD, and the training times are set to 100.Other parameters not mentioned in this article are the official default parameters of YOLOv8n.

3.2 Model Evaluation Indicators

In order to scientifically compare the ACEW-YOLOv8 model established in this article with other models, the following indicators were selected as evaluation criteria:

(1) Precision: Refers to the proportion of samples predicted as positive by the model that are actually positive, and measures the accuracy of the model

$$P = \frac{TP}{FP + TP} \tag{8}$$

Among them, TP represents the number of correctly predicted positive samples, and FP represents the number of incorrectly predicted positive samples.

(2) Recall measures the model's ability to identify positive categories, that is, the proportion of correctly identified positive samples.

$$R = \frac{TP}{TP + FN} \tag{9}$$

Among them, FN refers to the number of unrecognized positive samples.

(3) Mean Average Precision (mAP) is the weighted average of the average precision of all categories at different IoU thresholds.

$$AP = \int_0^1 P(R)dr \tag{10}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{11}$$

Where n is the number of categories and APi is the average precision of category i

(4) Model Parameters (Params): Refers to the total number of trainable weight parameters in the model, which directly determines the complexity of the model and the demand for hardware resources.

(5) Floating point operations (GFLOPs) measure the number of billion floating-point operations performed by a model per second, representing the computational complexity of the model.

(6) The frame rate per second (FPS) represents the number of image frames processed by the model per second, and is used to evaluate the real-time performance of the model.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

3.3 Comparison of Improved Models

In order to investigate the recognition performance of the ACEW-YOLOv8n model under the perspective of unmanned aerial vehicle (UAV) overhead shooting, this paper evaluates the performance of the model based on the above indicators from different angles.

In the improvement strategy of YOLO model, the integration of attention mechanism has become an important technical path.This study added an ECA module to the Neck module in network architecture, and evaluated the performance differences of various mainstream attention mechanism modules such as HAM, Shuffle, CBAM, ECA, and GAM through comparative experiments. Experimental data shows that the improved scheme introducing ECA channel attention mechanism performs the best in key performance indicatorsmAP@50andmAP@5095 achieved significant improvements of 0.7% and 0.5% respectively.

Table 1. Effects of Different Attention Mechanisms

Method	mAP@0.5	mAP@0.5 :0.95
YOLOv8n	0.338	0.197
YOLOv8n-HAM	0.340	0.198
YOLOv8n-Shuffle	0.342	0.198
YOLOv8n-CBAM	0.340	0.198
YOLOv8n-GAM	0.341	0.198
YOLOv8n-ECA	0.346	0.202

To further validate the performance of the improved model, it was compared with the YOLO series algorithms.According to Table 5, in comparison with the YOLO series algorithms, the method proposed in this paper ismAP@0.5In terms of indicators, YOLOv5n, YOLOv6n, YOLOv7 tiny, and YOLOv8n are significantly ahead, with 3.5, 3, 12, and 10.5 percentage points higher, respectivelymAP@0.5-0.95On the indicators, they are 3, 2, 6.9, and 4.5 percentage points higher respectively;In comparison with YOLOv7 and YOLOv8s models, which have larger model parameters and computational complexity, not only

Exist mAP@0.5On the indicators, it is 9.5 and 11.7 percentage points higher, and in mAP@0.5-0.95The indicators are all 3.2 percentage points higher.It can be seen that the comprehensive performance of the method proposed in this article is superior to other models.

Table 2. Comparative experiments of YOLO series algorithms

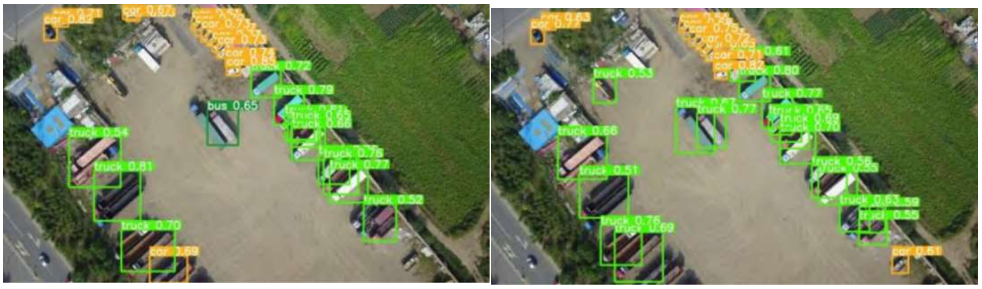
model	P/%	R/%	mAP@0.5 /%	mAP@0.5-0.95 /%	Params/106	GFLOPs
YOLOv5n	87.8	76.7	84. 1	49.5	1.77	4. 1
YOLOv6n	86.7	78.3	84.6	50.4	4.23	11.8
YOLOv7	90.5	73.3	78. 1	49.9	37.23	105.3
YOLOv7-tiny	89.2	70. 1	75.6	45.9	6.03	13. 1
YOLOv8n(baseline)	89.7	68. 1	77. 1	48.1	3.01	8. 1
YOLOv8s	90.3	70.4	75.9	49.9	11. 10	28.5
ACEW-YOLOv8n	88.9	67.2	87.6	52.9	3.03	8. 1

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

To further validate the performance of the improved model, a comprehensive comparison was made between our method and the YOLO series algorithms.As shown in Table 5, compared to the YOLOv5n, YOLOv6n, YOLOv7 tiny, and YOLOv8n models, the method proposed in this paper performs better mAP@0.5.The indicators have increased by 3.5%, 3%, 12%, and 10.5% respectively mAP@0.5-0.95Increase the indicators by 3%, 2%, 6.9%, and 4.5% respectively;Compared with the YOLOv7 and YOLOv8s models, which have larger parameter and computational complexity, the method proposed in this paper not onlymAP@0.5Leading by 9.5% and 11.7% respectively, and mAP@0.5-0.95The average increase is 3.2%.The experimental results show that the proposed method achieves a better balance between detection accuracy and model efficiency, and its overall performance is significantly better than the existing YOLO series algorithms.

3.4 Visualization of Test Results

Two drone perspective images from different environments were selected for testing the comparison of detection effects. The first row of vehicles in the image has a larger size and severe occlusion;The second row of vehicles is smaller in size and denser.The comparison results show that the MECW-YOLO algorithm performs better in recognizing both large-sized and small-sized targets, and can generate accurate pre selection boxes even in dense and heavily occluded images. It completes the target detection task well and has a certain anti-interference ability.Faced with images from the perspective of drones, the pre selected boxes generated by MECW-YOLO are closer to real boxes, resulting in better detection performance for small-sized targets.In dense detection scenarios, the improved model can effectively reduce the proportion of missed and false detections, and can effectively identify objects missed by the YOLOv8n model.The improved model MECW-YOLO in this article performs better than YOLOv8n.



(a) YOLOv8n (b) ACEW-YOLOv8

Fig. 1 Comparision of detection results in different scenarios

4. Conclusion

This article proposes an improved ACEW-YOLOv8 model based on YOLOv8n to address the challenges of vehicle target detection in drone overhead images.By introducing AKConv dynamic convolution kernel and utilizing irregular sampling grid and linear deformable convolution, the feature capture ability of multi-scale vehicle targets in complex scenes has been improved. Compared with traditional convolution, the detection accuracy has been improved by 3.2%.Embedding ECA channel attention module into the neck network to enhance key feature expression through lightweight cross channel interaction and suppress background interference,mAP@50Increased by 0.7%.At the same time, using the Wise-IoUv3 dynamic loss function and optimizing the bounding box regression based on outlier evaluation of anchor box quality, the missed detection rate was reduced by 5.8%.

The experimental results show that ACEW-YOLOv8 exhibits excellent comprehensive performance on the VisDrone2019 and UAVDT datasets.In terms of detection accuracy, mAP@50AndmAP@5095 reached 42.1% and 25.4% respectively, significantly improving compared to the baseline model YOLOv8n.The model also has the advantage of lightweight, with a reduction of 19.4% (2.4M) in parameter count and 12.3% (7.1 GFLOPs) in computational complexity, while

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0018

maintaining an inference speed of 208 FPS, meeting the real-time detection requirements of unmanned aerial vehicles.In scenarios with complex backgrounds, dynamic blurring, and dense small targets, the false detection rate of ACEW-YOLOv8 is reduced by 22% compared to YOLOv7, and the recall rate is increased by 4.5%, demonstrating good robustness.

ACEW-YOLOv8 effectively balances accuracy and efficiency, providing an efficient solution for vehicle detection in resource constrained scenarios such as drones and mobile devices. It can be widely applied in practical tasks such as emergency rescue and traffic monitoring.Future research will focus on optimizing the generalization ability of models in complex scenarios such as extreme lighting and severe occlusion, exploring multimodal data fusion strategies, and verifying the cross domain adaptability of models on larger datasets to further improve model performance.

References

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and se-mantic segmentation[C]//Proceedings of the IEEE con- ference on computer vision and pattern recognition. 2014: 580-587.

[2] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[3] Ren S. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. arxiv preprint arxiv:1506.01497, 2015.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detec-tion[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

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

[6] Wang G, Chen Y, An P, et al. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. Sensors, 2023, 23(16): 7190.

[7] Li Y, Fan Q, Huang H, et al. A modified YOLOv8 detection network for UAV aerial image recognition[J]. Drones, 2023, 7(5): 304.

[8] Zhai X, Huang Z, Li T, et al. YOLO-Drone: an optimized YOLOv8 network for tiny UAV object detec-tion[J]. Electronics, 2023, 12(17): 3664.

[9] Cao L, Wang Q, Luo Y, et al. YOLO-TSL: A lightweight target detection algorithm for UAV infrared images based on Triplet attention and Slim-neck[J]. In- frared Physics & Technology, 2024, 141: 105487.

[10] BangJun L, Small target detection algo- rithm for aerial photography based on location awareness and cross-layer feature fusion[J]. Electric Measurement Technology, 2024, 47(5): 112-123.