Lightweight Low-Altitude UAV Object Detection Based on Improved YOLOv5s

Haokai Zeng
College of Ordnance Science and Technology
Xi'an Technological University
Xi'an, 710021, China
E-mail: zhk790743337@outlook.com

Jing Li
College of Electronic Information Engineering,
College of Ordnance Science and Technology
Xi'an Technological University
Xi'an, 710021, China
E-mail: gem99li@163.com

Liping Qu
College of Ordnance Science and Technology
Xi'an Technological University
Xi'an, 710021, China
E-mail: quliping@qq.com

***Abstract*—In the context of rapid developments in drone technology, the significance of recognizing and detecting low-altitude unmanned aerial vehicles (UAVs) has grown. Although conventional algorithmic enhancements have increased the detection rate of low-altitude UAV targets, they tend to neglect the intricate nature and computational demands of the algorithms. This paper introduces ATD-YOLO, an enhanced target detection model based on the YOLOv5s architecture, aimed at tackling this issue. Firstly, a realistic low-altitude UAV dataset is fashioned by amalgamating various publicly available datasets. Secondly, a C3F module grounded in FasterNet, incorporating Partial Convolution (PConv), is introduced to decrease model parameters while upholding detection accuracy. Furthermore, the backbone network incorporates an Efficient Multi-Scale Attention (EMA) module to extract essential image information while filtering out irrelevant details, facilitating adaptive feature fusion. Additionally, the universal upsampling operator CARAFE (Content-aware reassembly of features) is utilized instead of nearest-neighbor upsampling. This enhancement boosts the performance of the feature pyramid network by expanding the receptive field for data feature fusion. Lastly, the Slim-Neck network is introduced to fine-tune the feature fusion network, thereby reducing the model's floating-point calculations and parameters. Experimental findings demonstrate that the improved ATD-YOLO model achieves an accuracy of 92.8%, with a 31.4% decrease in parameters and a 28.7% decrease in floating-point calculations compared to the original model. The detection speed reaches 75.37 frames per second (FPS). These experiments affirm that the proposed enhancement method meets the deployment requirements for low computational power while maintaining high precision.**

Keywords-Lightweight; Small Object; UAV Detection

I. INTRODUCTION

Small UAVs are aircraft known for their diminutive size, cost-effectiveness, and ability to fly at low altitudes. Due to their compactness and ease of operation, small UAVs have emerged prominently on modern military battlegrounds, garnering significant favor. They demonstrate exceptional performance in tasks such as reconnaissance, surveillance, communication, and target identification. However, the low-altitude, slow-flight attributes of small UAVs render them challenging to effectively counter using conventional detection methods, thus presenting novel challenges for military defense. Consequently, detecting UAV targets in flight has become a pivotal approach to addressing this issue[1].

Utilizing image and video analysis, employing computer vision algorithms for real-time UAV detection and tracking stands as the most promising method for UAV detection. In comparison to standard radar-based methodologies, this system offers a myriad of advantages, encompassing enhanced accuracy and reduced costs [2].

To date, the majority of detection tasks have been predominantly conducted through the utilization of deep learning methodologies for

DOI: 10.2478/ijanmc-2024-0009

feature extraction. This method mainly consists of two types of algorithms: two-stage and single-stage object detection. The former involves initially delineating the regions of interest before determining target position and class information. Representative algorithms include R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], and Mask R-CNN [6]. The latter directly ascertain target position and class information without the need for separately identifying regions of interest. Typical algorithms include YOLO [7-9] and SSD [10]. Given the high maneuverability of low-altitude UAVs, capable of swiftly altering direction, moving at high velocities, and executing diverse flight maneuvers, numerous scholars opt for the YOLOv5 algorithm and its enhancements to execute UAV target detection tasks.

Lu et al. [11] introduced an improved YOLOv5s-based algorithm for small rotary-wing UAV target detection, demonstrating enhanced accuracy and feature extraction capabilities, it experiences a certain decrease in detection speed. Yang et al. [12] developed a real-time detection algorithm, named GCB-YOLOv5s, for low-altitude UAVs using machine vision detection techniques. While this algorithm boosts detection speed, it also leads to a slight decline in detection accuracy. Bao et al. [13] presented a real-time detection method for micro UAVs based on YOLOv5. Although the algorithm demonstrates commendable real-time performance for UAV targets at low altitudes, its effectiveness in detecting UAV targets in distant scenes is limited, and its robustness is relatively poor.

In summary, existing algorithms have improved the detection accuracy of low-altitude UAV targets but have overlooked the complexity and computational burden of the algorithms. Hence, the engineering challenge at hand is: how to enhance the algorithm's detection efficiency for UAV targets while preserving detection accuracy, employing lightweight design principles. Consequently, this paper suggests a lightweight detection algorithm for low-altitude UAVs, coined ATD-YOLO, and based on an enhanced version of YOLOv5s. The key enhancements of this algorithm are as follows:

1. By merging multiple publicly available datasets, a relatively comprehensive UAV target dataset is constructed.2. Based on the lightweight model FasterNet [14], the paper proposes a lightweight module called C3F to substitute the C3 module in the input feature extraction network. This substitution substantially reduces the number of parameters and floating-point calculations, thereby achieving lightweight effects on the overall network. 2. A more optimal upsampling method, CARAFE [15] is employed to increase the receptive field, enhancing feature sharpness post traditional upsampling. 3. EMA [16] is integrated into the backbone network to extract vital image information while filtering out irrelevant details, thus enabling adaptive feature fusion and enhancing detection accuracy. 4. Slim-Neck [17] is introduced into the neck part of the network, replacing Conv layers and C3 layers with lightweight convolutional neural networks GSConv and VOVGSCSP. This further reduces the computational workload and parameter complexity of the model, thereby improving its inference speed without sacrificing detection accuracy.

The improved ATD-YOLO network structure, shown in Figure 1, maintains detection accuracy while adopting a lightweight design, meeting the demands for real-time operation for UAV target detection. The enhanced detection model better accommodates the computational constraints of UAV detection devices, providing a research solution for lightweight improvements in UAV target detection.

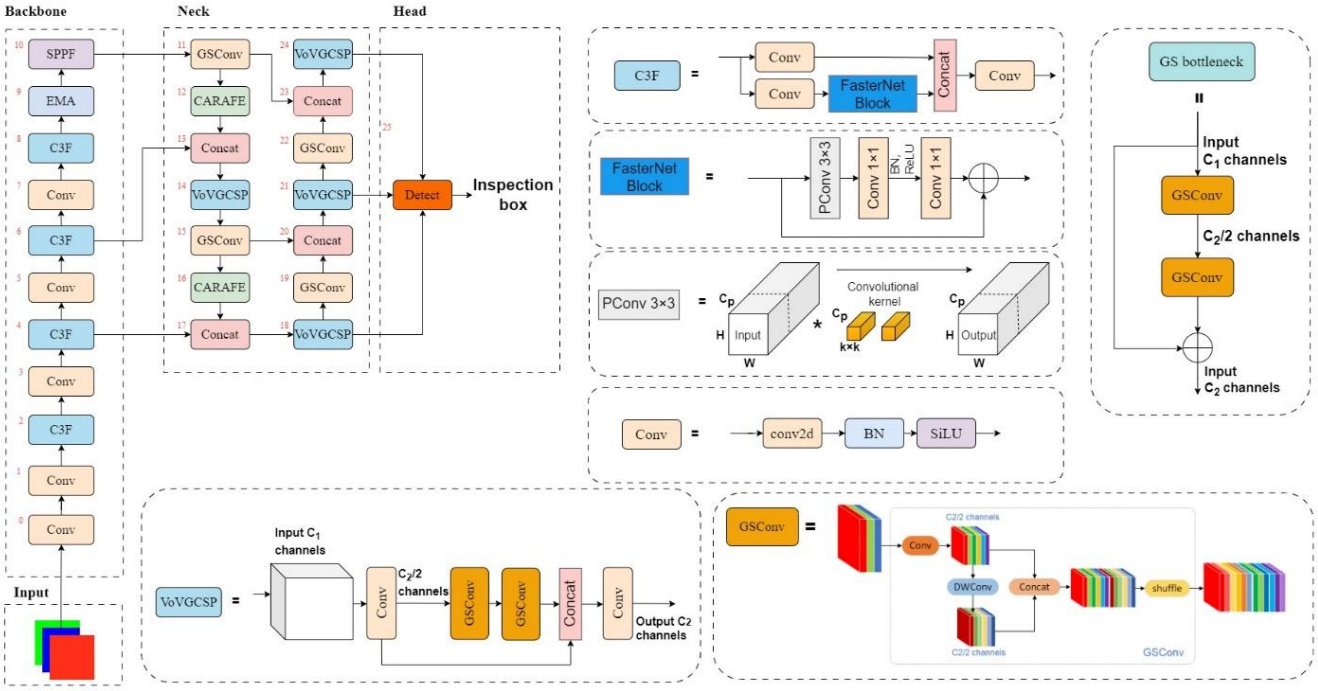


Figure 1. ATD-YOLO Network Structure

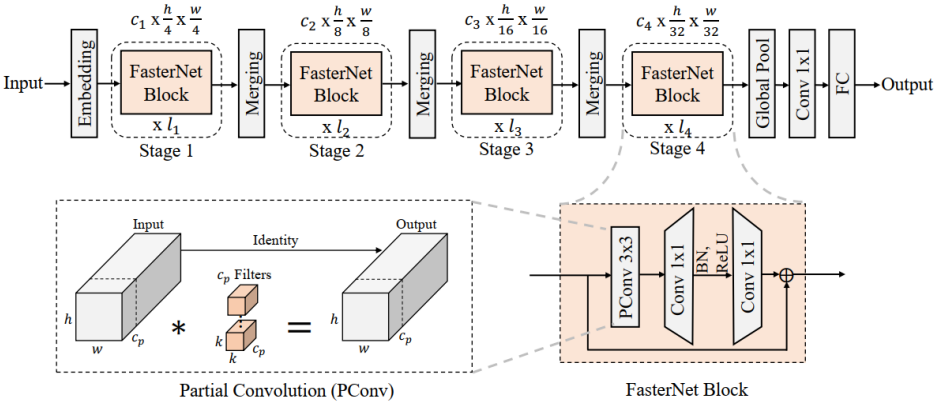


Figure 2. Framework of image measuring system [14]

II. Algorithm Improvement And Optimization

A. C3F Lightweight Feature Extraction Module

Addressing redundant computation, Chen et al. [14] introduced Partial Convolution (PConv), which reduces memory access while optimizing the parameter problem caused by redundant computation, greatly improving the ability to capture spatial features. FasterNet is constructed with PConv and 1x1 convolutional structures. In Figure 2, h, w, and k represent the height, width, and kernel size of the feature map, respectively, while Cp indicates the number of channels in conventional convolution.

PConv only uses general Conv to achieve spatial feature acquisition on some input channels, while maintaining the remaining channels unchanged. Calculate by considering the first or last consecutive Cp channel as a representation of the entire feature map. Ensure its generality while maintaining the same number of input and output feature map channels. The FLOPs of PConv are h × w × k2 × C2p, which only accounts for 1/16 of the general Conv.

The C3 module in the YOLOv5 network is pivotal for increasing network depth and receptive field, enhancing feature extraction. Initially, it included Conv1, Conv2, Conv3 modules, and one or more Bottleneck modules. While this design enriches the learning capabilities of the C3 module, it also adds to the computational load and model complexity. Thus, this paper introduces a lightweight feature extraction module, C3F, inspired by the FasterNet module concept.

Figure 3 illustrates the structural diagram of the C3F feature extraction module. Here, h, w, and k signify the height, width, and convolution kernel size of the feature map, while Cp represents the number of conventional convolution channels.This module introduces partial convolution and replaces BattleNet with FasterNet Block in the C3 module, reducing computational redundancy and memory access while maintaining the speed and efficiency of feature extraction.

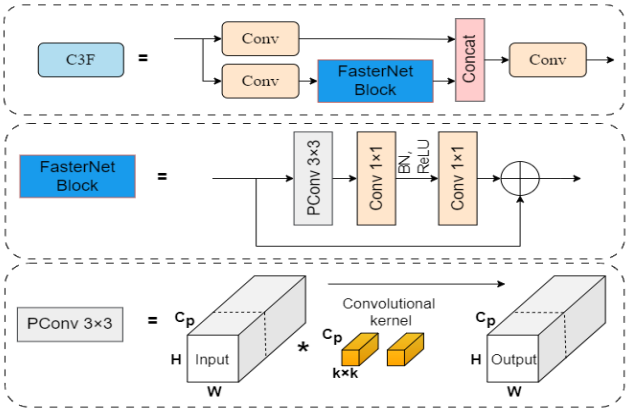


Figure 3. C3F structural schematic diagram

B. CARAFE upsampling module

Object detection models often employ nearest neighbor or bilinear interpolation for feature map upsampling. While adaptive upsampling uses methods such as deconvolution. However, these traditional methods have certain shortcomings in accurately reconstructing target detail information, which can easily lead to partial information loss of small targets, thereby affecting detection accuracy. In contrast, CARAFE improves the quality of upsampled features by recombining content aware features to make the upsampling kernel semantically relevant to the feature map. The CARAFE operator can better preserve and recover feature information details, so this article chooses to use the CARAFE operator for upsampling to enhance regional sensitivity and generate more accurate high-resolution feature maps.

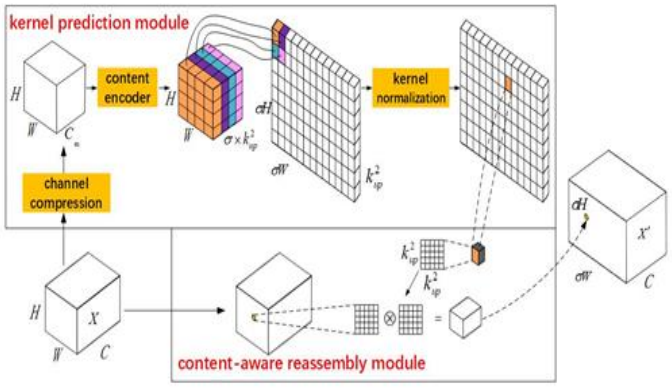


Figure 4. CARAFE upsampling calculation flowchart

C. Multi scale attention mechanism module

In various computer vision tasks, the significant effectiveness of channel or spatial attention mechanisms in generating clearer feature representations has been demonstrated. However, modeling cross channel relationships through channel dimensionality reduction may have side effects on extracting deep visual representations.

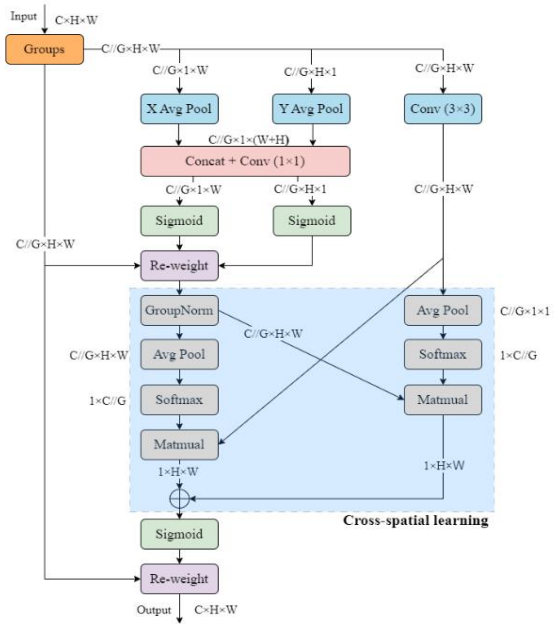


Figure 5. EMA Attention Mechanism

EMA without dimensionality reduction [16] aims to preserve channel information while minimizing computational overhead. This is achieved by reshaping some channels into batch dimensions and grouping channel dimensions into multiple sub-features to evenly distribute spatial semantic features. EMA learns effective channel descriptions in convolutional operations without reducing channel dimensionality, enhancing pixel-level attention for advanced feature maps. This lightweight and flexible EMA attention model serves as a core module applicable to lightweight networks.

Figure 5 outlines the EMA attention module, comprising Feature Grouping, Parallel Subnetworks, and Cross spatial learning. EMA divides the input feature map into multiple sub-features based on channel dimensions. It employs two parallel subnetworks: one with 1x1 branches and the other with 3x3 branches. The 1x1 branch encodes channel information using global average pooling operations, while the 3x3 branch captures local cross-channel interaction. EMA then fuses the output feature maps of the two subnetworks using cross-space learning, resulting in an attention weight map of the same size as the input feature map to enhance its expressive power.

D. Lightweight fusion stage

Large deep learning models are difficult to deploy on industrial embedded devices. Many lightweight networks use a large number of depthwise separable convolutions, and even if the C3 of the backbone network is replaced with lighter modules, there are still a large number of 1x1 convolution operations, making it difficult to achieve sufficient accuracy. This dense convolution operation actually consumes more resources, and even with channel shuffling, the effect is still poor. Therefore, this article embeds the Slim Neck [17] network is integrated into the feature fusion stage, incorporating the GSConv module and the VOVGSCSP module.

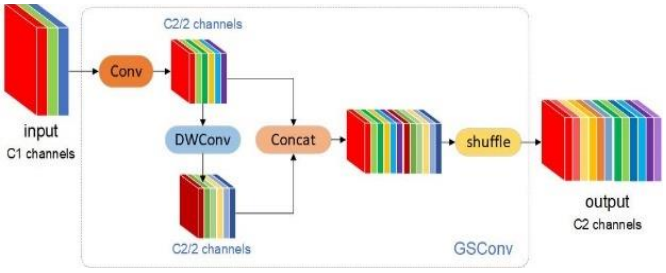


Figure 6. GSConv Module

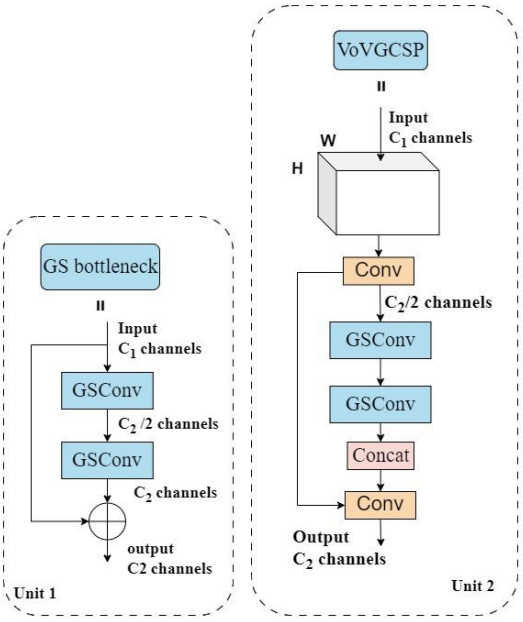


Figure 7. VoVGCSP Module

The GSConv module consists of the Conv, DWConv, Concat, and Channel Mixing sub-modules, illustrated in Figure 6. Its operational procedure is as follows: Initially, the input feature map, with C1 channels, undergoes standard convolution to produce a feature map with C2/2 channels. Subsequently, depthwise separable convolution generates another feature map with C2/2 channels. These two feature maps are concatenated to form a unified feature map with C2 channels. Finally, the channel mixing operation adjusts the output characteristics to the desired channel count. Through this approach, the GSConv module combines depthwise separable convolution and standard convolution to reduce computational complexity and improve overall recognition accuracy by addressing limitations in feature extraction and fusion.

The integration of GSConv convolution aims to simplify model complexity. To enhance model inference speed without sacrificing accuracy, we introduced the VOVGSCSP module, as depicted in Figure 7. In Figure 7, Unit 1 illustrates the bottleneck unit structure of VOVGSCSP, while Unit 2 showcases a cross-stage VOVGSCSP module employing a single aggregation method.

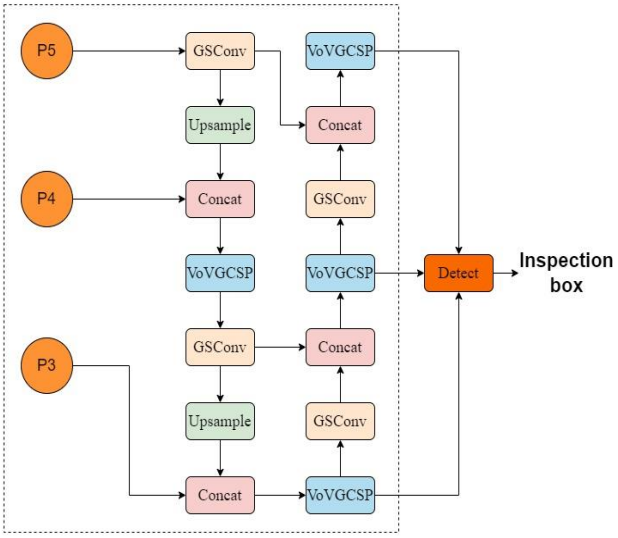


Figure 8. The positions of GSConv and VOVGCSP modules

III. Experimental Testing and Result Analysis

A. Experimental Design and Parameter Setting

In the identical operational setting and employing the UAV dataset, experimental tests are conducted to compare the enhanced ATD-YOLO algorithm model with other algorithm models. The experimental environment configuration is outlined in Table 1. This comparison seeks to validate the improved algorithm's efficacy in detecting low-altitude UAV targets.

Throughout the training phase, a combination of the cosine annealing learning rate decay method and the SGD algorithm is utilized. The training regimen spans 600 epochs with a batch size of 16 and a momentum of 0.937. Mosaic data augmentation is employed to enrich the backgrounds of images by randomly scaling, cropping, and arranging four training images in a mosaic pattern. This augmentation technique enhances the accuracy and robustness of small object detection.

TABLE I. Experimental Setup Configuration

Name	Environment Configuration
System Environment	Ubuntu 22.04
CPU	AMD Ryzen 9 5950X
GPU	RTX 4060 Ti 16GB
Deep Learning Framework	Pytorch 1.13.1
IDE	CUDA 11.7

B. Building a Relatively Comprehensive Dataset

In this study, extensive data from diverse sources and papers was reviewed and collected. Utilizing this data, we constructed a tailored dataset called "Anti-Mini Drone" for our research purposes. The Det-Fly dataset [18] addresses the lack of drone data from a single perspective by directly collecting images of target drones in the air, including various postures such as upward, downward, and forward views. However, this dataset only contains one type of drone, limiting its generality in detecting other types of drones. The Drone-vs-Bird dataset [19] not only covers rich drone and environmental data but also includes some bird data. This poses challenges when drones resemble birds in appearance, especially during long-distance observations. However, the drawback of this dataset is its inability to meet the detection requirements of other types of drones. The Real World dataset [20] contains various types of drones and environments sourced from YouTube videos, but the image resolution is low. Most of the data is captured from a forward and upward perspective, which implies certain limitations in drone detection from a downward view. The Multi-view drone tracking dataset [21] records drone flight trajectories from different angles using multiple consumer-grade cameras, but the environmental capture is relatively homogeneous. The DUT anti-UAV dataset [22] consists of both a detection dataset comprising 10,000 images and a tracking dataset containing 20 videos. However, it's worth noting that the distribution of target dimensions within the dataset is uneven. The Anti-UAV dataset [23] includes visible light and infrared data, but the problem is that the shooting environment is singular, suitable only for research on multi-modal

fusion tracking and the alignment between infrared and visible light cameras is not perfect in time and space.

Overall, the above drone datasets each have their own advantages and disadvantages, often overcoming only one or two difficulties in constructing drone datasets. By merging these six different datasets, a low-altitude drone dataset that meets the requirements of this study was constructed, making it more reflective of real outdoor flight scenarios for low-altitude drones. Table 2 illustrates the distribution of images.

TABLE II. ORIGIN OF THE DATASET AND QUANTITY OF IMAGES

Dataset	Number of Images
Det-Fly	3893
Drone-vs-Bird	3959
Real World	1525
Multi-view drone tracking	3447
DUT anti-UAV	3639
Anti-UAV	2767

In the Anti-Mini Drone dataset, Figure 9 demonstrates a notable prevalence of small targets. The majority of targets in the dataset exhibit aspect ratios less than 0.1 times the original image dimensions. This distribution aligns with the relative sizes of objects commonly encountered in real outdoor scenarios during low-altitude unmanned aerial vehicle (UAV) flights.

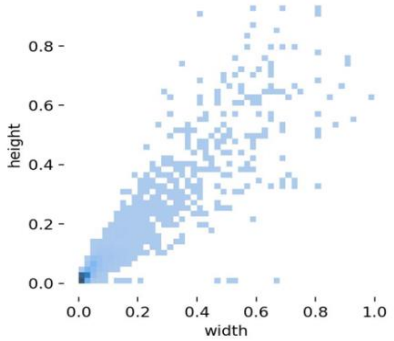


Figure 9. Length and Width Distribution Chart of the Anti-Mini Drone Dataset

C. Experimental evaluation metrics

To evaluate the enhanced ATD-YOLO algorithm, metrics such as parameter count, floating-point operations (GFLOPs), average precision (AP), and frames per second (FPS) are selected. Since the study focuses on detecting drones across different categories, mean average precision (mAP) and AP values are considered equivalent. Given the predominance of small objects, the mAP.5 criterion is adopted for evaluation to reflect the model's performance and speed accurately. Precision measures the proportion of correctly detected objects, while recall assesses the proportion of correctly predicted objects among all true objects:

$$\text{Precision} = \frac{\text{TP}}{TP + FP} \tag{1}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{2}$$

FN represents false positive samples predicted by the model, TP denotes true positive samples predicted correctly, and FP stands for false negative samples. The average precision (AP) reflects the detection accuracy for a single class of targets, usually calculated by integrating the Precision-Recall (P-R) graph:

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

Detection speed is frequently measured in FPS (Frames Per Second), indicating the number of images processed by the object detection network per second. A higher FPS value signifies faster processing speed. The expression for FPS is given by Formula 4:

$$FPS = \frac{FrameNum}{ElapsedTime} \tag{4}$$

(a) Sky Background (b) Treetops Background (c) Dusk Background (d) Evening Background

Figure 10. Samples of simple background from Anti-Mini Drone

(a) Mountainous background (b) Wooded background (c) Similar colors (d) Semi-obscured

Figure 11. Samples of complex background from Anti-Mini Drone

D. Experimental Results Analysis

a)Comparative Experiment on Lightweight Modules: Experimental analysis confirmed the effectiveness of the C3F module. Replacing the C3 module in YOLOv5s with C3F increased mAP.5 by 0.1%, while reducing parameters and floating-point operations by 9.7% and 12.6%, respectively, and improving FPS by 6.26. Substituting with C2f increased mAP.5 accuracy by 0.9%, but raised parameters and floating-point operations by 17.6% and 22.7%, respectively, while decreasing FPS by 13.41. Replacing with C2f-Faster decreased mAP.5 by 0.3%, with parameters and floating-point operations decreasing by 8.2% and 6.1%, respectively, and FPS increasing by 8.78. These results validate the effectiveness of the C3F module in achieving high accuracy with minimal algorithmic overhead.

TABLE III. CONTRAST EXPERIMENT OF ATTENTION MODULE

Module	mAP.5/%	GFLOP /G	Params/106	FPS
C3	92.2	15.8	7.01	68.79
C3F	92.3	13.8	6.33	75.05
C2f	93.1	19.4	8.25	55.38
C2f-Faster	91.9	14.5	6.58	77.57

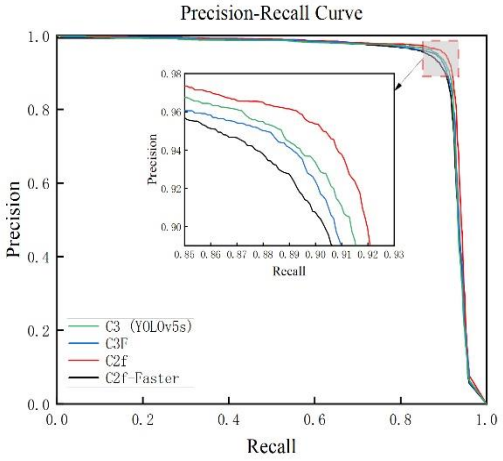


Figure 12. PR curves for various feature extraction modules (IOU=0.5)

Figure 12 illustrates the PR curves of various feature extraction modules at IOU = 0.5. It also demonstrates that the improved algorithm's Precision and Recall with the C3F module are slightly lower than those with the original C3 module. However, considering the improvement in detection accuracy, parameter count, floating-point operations, and frame rate, the improved model still exhibits superiority in target localization regression.

b)Comparative Experiment on Attention Mechanism Modules: In response to the observed increase in model parameter count, floating-point operations, and FPS after adding EMA, various attention mechanisms were replaced at the

original position for comparative experiments. Table 4 presents the experimental results. The improved algorithm with EMA sacrifices some performance compared to other attention mechanism algorithms but achieves better detection accuracy.

TABLE IV. CONTRAST EXPERIMENT OF ATTENTION MODULE

Module	mAP.5/%	GFLOP /G	Params/106	FPS
SE[24]	91.8	13.8	6.37	74.93
ECA[25]	92.2	13.8	6.34	74.37
CBAM[26]	92.3	13.8	6.37	72.63
CA[27]	91.2	13.8	6.36	72.87
EMA	92.7	14.1	6.38	69.83

Figure 13 illustrates the PR curves comparing various attention mechanisms at IOU = 0.5. It also demonstrates that the EMA attention mechanism outperforms other attention mechanisms in both Precision and Recall.

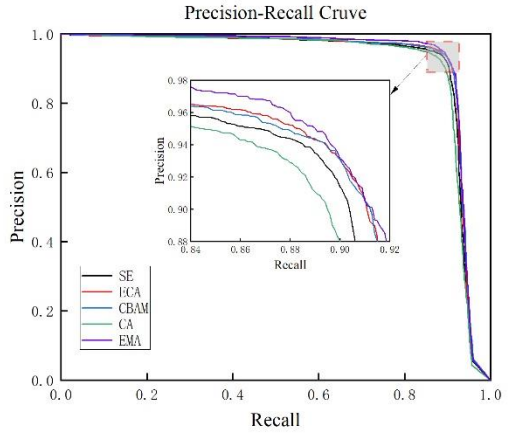


Figure 13. PR curves for various feature extraction modules (IOU=0.5)

c)Ablation Experiment: Ablation experiments were conducted on the Anti-Mini Drone dataset to test different model configurations, including variations in the C3F lightweight feature extraction module, EMA attention mechanism module, CARAFE upsampling module, and Slim-Neck module. Results from these experiments are summarized in Table 5.In the first row, the initial YOLOv5s model achieved a detection accuracy of 92.2%. By replacing the C3 module with the C3F module, the parameter count decreased by about 17.2%, while floating-point operations increased by 5.6%. The mAP.5 increased by 0.1 percentage points, and the FPS improved by 6.26, indicating that adopting FasterNet to enhance the C3 module effectively reduces model parameters while maintaining detection capabilities.In the third row, adding the EMA attention mechanism module increased mAP.5 by 0.4% compared to the second row. However, parameter count and floating-point operations increased by 1.1% and 2.1%, respectively, while FPS decreased by 5.22.In the fourth row, after introducing the CARAFE upsampling module, parameter count increased by about 0.8%, while floating-point operations increased by 0.7%. The mAP.5 increased by 0.4 percentage points, while FPS decreased by 1.98.In the fifth row, embedding the Slim-Neck network resulted in a decrease of 18.28% in parameter count and 21.98% in floating-point operations compared to the fourth row. Despite a slight decrease of 0.3% in mAP.5, FPS increased by 7.5. Compared to the initial model in the first row, parameter count and floating-point operations decreased by 25.39% and 30.37%, respectively, while mAP.5 and FPS increased by 0.5% and 6.56%, respectively.

TABLE V. RESULTS OF ABLATION EXPERIMENTS

YOLOv5s	C3F	EMA	CARFE	Slim-Neck	Params/106	GFLOP/G	mAP.5/%	FPS
√					7.01	15.8	92.2	68.79
√	√				6.33	13.8	92.3	75.05
√	√	√			6.38	14.1	92.7	69.83
√	√	√	√		6.40	14.1	93.1	67.85
√	√	√	√	√	5.23	11.0	92.8	75.35

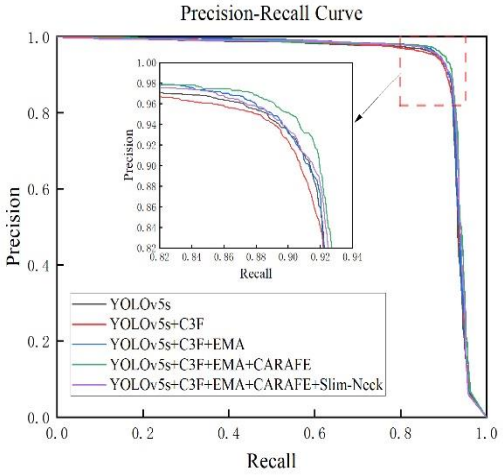


Figure 14. Model PR curve (IOU=0.5)

The findings indicate improved detection accuracy with reduced parameter count and floating-point operations, confirming the effectiveness of the enhancement strategies. Figure 14 displays PR curves of the target detection algorithm at an IOU of 0.5, showing that the improved algorithm outperforms the original in both Precision and Recall, highlighting its superior performance in target localization regression.

d)Comparative Experiment on Mainstream Algorithms: Experimental analysis compared the effectiveness of the C3F, C2f, and C2f-Faster modules in feature extraction networks. Results in Table 3 show that replacing C3 with C3F increased mAP.5 by 0.1%, with reduced parameters and floating-point operations and increased FPS. Substituting with C2f improved mAP.5 by 0.9% but increased parameters and floating-point operations while decreasing FPS. Replacing with C2f-Faster resulted in a 0.3% decrease in mAP.5, with reduced parameters and floating-point operations and increased FPS. These results confirm the effectiveness of the C3F module in achieving high accuracy with minimal overhead, demonstrating its superiority in lightweight feature extraction.

TABLE VI. MAINSTREAM ALGORITHM COMPARATIVE EXPERIMENT RESULTS

Module	Params/106	GFLOP/G	AP.5/%	FPS
YOLOv3 Tiny	8.66	12.9	79.1	166.67
YOLOv5s	7.01	15.9	92.2	68.79
YOLOv7 Tiny	6.01	13.2	88.4	63.30
YOLOv8s	11.12	28.4	89.0	109.89
ATD-YOLO	5.23	11.0	92.8	75.35

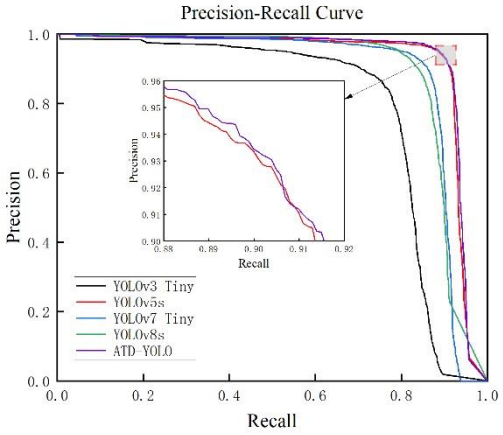


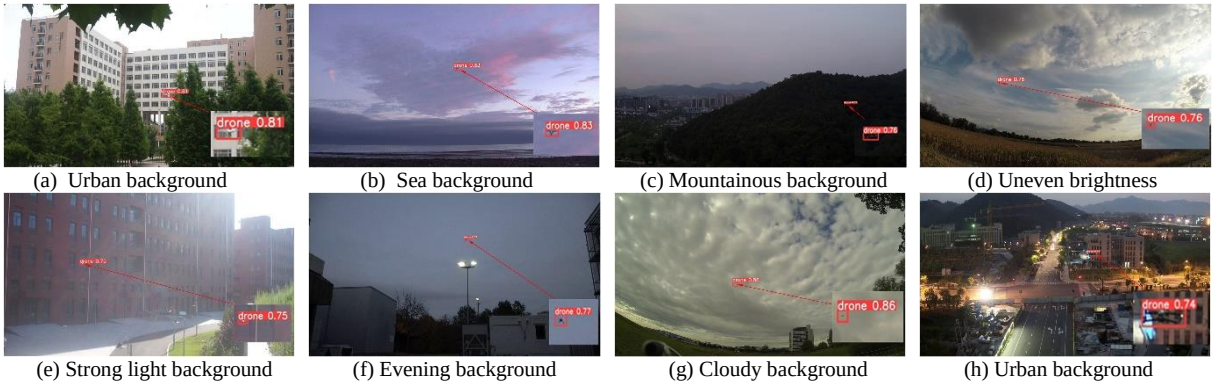
Figure 15. PR curves of mainstream algorithms on the test set (IOU=0.5)

Figure 15 displays PR curves for various algorithms on the test set at IOU = 0.5, indicating the improved algorithm's superiority in both Precision and Recall.

E. Analysis of Comparative Experiment Results

a) Comparative Experiment: The improved algorithm's detection results in various scenarios are intuitively and clearly demonstrated in Figure 12, indicating its superiority over the original model in drone object detection across different scenes, primarily reflected in confidence and detection outcomes. Specific scenarios include background with buildings (Figure 16 (a)), flying over the sea (Figure 16 (b)), flying in mountainous areas (Figure 16 (c)), flying under uneven brightness conditions (Figure 16 (d)), flying in strong sunlight conditions (Figure 16 (e)), flying in the evening (Figure 16 (f)), flying with cloud backgrounds (Figure 16 (g)), and flying in urban backgrounds (Figure 16 (h)).In the background with buildings, even with interference

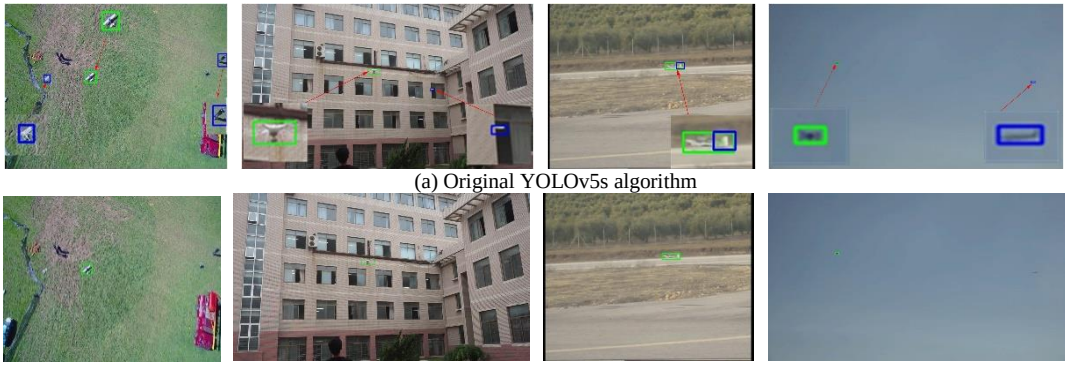
such as trees and buildings, where the background is complex and the drone is similar in height to the background, the improved algorithm can successfully detect the target. In other scenarios such as flying over the sea, in mountainous areas, under uneven lighting, in strong sunlight, in the evening, with cloud backgrounds, and in urban backgrounds, the improved algorithm also performs admirably, successfully detecting drone targets without missing or false detections. This firmly establishes the effectiveness of the improved algorithm in complex scenarios.



(a) Urban background (b) Sea background (c) Mountainous background (d) Uneven brightness

(e) Strong light background (f) Evening background (g) Cloudy background (h) Urban background

Figure 16. Object detection outcomes in diverse scenarios



(a) Original YOLOv5s algorithm

(b) ATD-YOLO algorithm

Figure 17. Object detection outcomes in a consistent scenario

b) Comparative Study: We compared the original YOLOv5s algorithm with the enhanced ATD-YOLO algorithm on a test dataset. In Figure 17, green boxes show correct identifications, while blue boxes indicate misidentifications. The original algorithm sometimes misidentified pedestrians and debris as targets when a drone was on the lawn. However, the improved algorithm performed better, correctly identifying targets and avoiding mistakes like thinking reflections were drones or confusing a drone's tail with the whole drone. Overall, the improved algorithm is better at detecting drones, making fewer mistakes while still being efficient.

IV. Conclusions

In addressing the challenge of improving the simultaneous accuracy and detection efficiency of low-altitude UAV target detection algorithms, this paper presents an enhanced algorithm, ATD-YLO, built upon YOLOv5s. This algorithm successfully achieves lightweight target detection, aiming to maintain detection precision and efficiency in low-altitude UAV detection tasks under limited hardware resource platforms.

ATD-YOLO introduces several innovations to improve performance. It includes PConv, a new convolutional layer, and C3F, a lightweight feature

extraction module inspired by FasterNet. C3F replaces the original C3 module, reducing parameters and computations while maintaining recognition accuracy. EMA, an attention mechanism module, is also integrated to extract key information from images while ignoring irrelevant data, enhancing detection accuracy. Furthermore, the introduction of CARAFE, a generic upsampling module, increases the receptive field for feature fusion, and Slim-Neck, a lightweight network, further promotes network efficiency.

The effectiveness of the proposed approach was validated by training and validating the improved ATD-YOLO algorithm on the Anti Mini Drone dataset. Experimental results revealed an accuracy increase from 92.2% to 92.8% compared to the initial algorithm. Furthermore, the improved algorithm reduced parameter count and floating-point computations by 31.4% and 28.9%, respectively, while achieving a detection speed of 75.35 FPS. The improved algorithm outperforms YOLOv3 Tiny, YOLOv7 Tiny, and YOLOv8s in recognition accuracy by 13.7%, 4.4%, and 3.8%, respectively, with model parameter counts of 60.39%, 87.02%, and 47.03% of theirs, and floating-point computations of 85.27%, 83.33%, and 38.73% of theirs, respectively. The FPS is 12.35 higher than that of YOLOv7 Tiny, but only 45.21% and 68.56% of YOLOv3 Tiny and YOLOv8s, respectively. Therefore, ATD-YOLO exhibits promising performance and meets the lightweight detection requirements for UAVs. In the next phase of research, efforts will focus on dataset expansion to include more categories such as birds in flight and other airborne objects, as well as improving network detection speed.

REFERENCES

[1] Zhao F, Zhao C, Guo J.Visual perception-based anti-drone technology: Development dynamics and trend [J]. National Defense Technology, 44(05), 35-45. DOI: 10.13943/j.issn1671-4547.2023.05.05.

[2] Oh H M, Lee H, Kim M Y .Comparing Convolutional Neural Network(CNN) models for machine learning-based drone and bird classification of anti-drone system [C]//2019 19th International Conference on Control, Automation and Systems (ICCAS).2019. DOI:10.23919/ICCAS47443. 2019. 8971699.

[3] 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.

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

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

[6] He K, Gkioxari G, Dollár P, et al. Mask r-cnn [C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

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

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement [J]. arXiv preprint arXiv:1804.02767, 2018.

[10] 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.

[11] LU Q, YU Y Q, XU D M, et al. Improved YOLOv5 Small Drones Target [J]. Computer Science, 2023, 50(S2): 212-219.

[12] YANG H Y, RONG Y S, JIAN Y H, et al. GCB-YOLOv5s algorithm for real-time detection for a low altitude UAV [J]. Journal of Ordnance Equipment Engineering, 2023, 44(07): 1-8.

[13] BAO W Q, XIE L Q, XU C, et al. A Real-time detection method of micro UAV based on YOLOv5 [J]. Journal of Ordnance Equipment Engineering, 2022, 43(05): 232-237.

[14] Chen J, Kao S, He H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12021-12031.

[15] Wang J, Chen K, Xu R, et al. Carafe: Content-aware reassembly of features [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3007-3016.

[16] Ouyang D, He S, Zhang G, et al. Efficient Multi-Scale Attention Module with Cross-Spatial Learning [C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[17] Li H, Li J, Wei H, et al. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles [J]. arXiv preprint arXiv: 2206.02424, 2022.

[18] Zheng Y, Chen Z, Lv D, et al. Air-to-Air Visual Detection of Micro-UAVs: An Experimental Evaluation of Deep Learning [J].IEEE Robotics and Automation Letters, 2021, PP(99): 1-1. DOI: 10.1109/LRA. 2021. 3056059.

[19] Coluccia A, Fascista A, Schumann A, et al. Drone vs. Bird Detection: Deep Learning Algorithms and Results from a Grand Challenge [J]. Sensors, 2021, 21(8): 2824. DOI:10.3390/s21082824.

[20] Pawelczyk M L, Wojtyra M .Real World Object Detection Dataset for Quadcopter Unmanned Aerial Vehicle Detection [J]. IEEE Access, 8:174394-174409 [2023-10-12]. DOI: 10.1109/ACCESS.2020. 3026192.

[21] Li J, Murray J, Ismaili D, et al. Reconstruction of 3D flight trajectories from ad-hoc camera networks [C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 1621-1628.

[22] J. Zhao, J. Zhang, D. Li and D. Wang, "Vision-Based Anti-UAV Detection and Tracking [J]. IEEE Transactions on Intelligent Transportation Systems, Dec. 2022, DOI: 10.1109/TITS. 2022.3177627.

[23] Jiang Nan, Wang Kuiran, Peng Xiaoke, et al. Anti-UAV: A large multi-modal benchmark for UAV tracking [J]. arXiv preprint arXiv, 2021. 2101(2), 1-13.

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[25] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020.

[26] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[27] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13708-13717.