electronics

MDPI

Article

YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection

Xianxu Zhai [1,2], Zhihua Huang [1,2,*], Tao Li [1,2], Hanzheng Liu [1,2] and Siyuan Wang [1,2]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi 830049, China; 107552103582@stu.xju.edu.cn (X.Z.)
[2] Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830049, China
* Correspondence: zhhuang@xju.edu.cn

Highlights:

What are the main findings?

- An improvement upon the state-of-the-art YOLOv8 model, proposing a high-performance and highly generalizable model for detecting tiny UAV targets.

What is the implication of the main finding?

- Addressing the small size characteristics of UAV targets, a high-resolution detection branch is added to the detection head to enhance the model's ability to detect tiny targets. Simultaneously, prediction and the related feature extraction and fusion layers for large targets are pruned, reducing network redundancy and lowering the model's parameter count.
- Improving multi-scale feature extraction, using SPD-Conv instead of Conv to extract multi-scale features, better retaining the features of tiny targets, and reducing the probability of UAV miss detection. Additionally, the multi-scale fusion module incorporates the GAM attention mechanism to enhance the fusion of target features and reduce the probability of false detections. The combined use of SPD-Conv and GAM strengthens the model's ability to detect tiny targets.

Abstract: With the widespread use of UAVs in commercial and industrial applications, UAV detection is receiving increasing attention in areas such as public safety. As a result, object detection techniques for UAVs are also developing rapidly. However, the small size of drones, complex airspace backgrounds, and changing light conditions still pose significant challenges for research in this area. Based on the above problems, this paper proposes a tiny UAV detection method based on the optimized YOLOv8. First, in the detection head component, a high-resolution detection head is added to improve the device's detection capability for small targets, while the large target detection head and redundant network layers are cut off to effectively reduce the number of network parameters and improve the detection speed of UAV; second, in the feature extraction stage, SPD-Conv is used to extract multi-scale features instead of Conv to reduce the loss of fine-grained information and enhance the model's feature extraction capability for small targets. Finally, the GAM attention mechanism is introduced in the neck to enhance the model's fusion of target features and improve the model's overall performance in detecting UAVs. Relative to the baseline model, our method improves performance by 11.9%, 15.2%, and 9% in terms of P (precision), R (recall), and mAP (mean average precision), respectively. Meanwhile, it reduces the number of parameters and model size by 59.9% and 57.9%, respectively. In addition, our method demonstrates clear advantages in comparison experiments and self-built dataset experiments and is more suitable for engineering deployment and the practical applications of UAV object detection systems.

Keywords: UAV; object detection; YOLOv8; deep learning

Citation: Zhai, X.; Huang, Z.; Li, T.; Liu, H.; Wang, S. YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection. *Electronics* **2023**, *12*, 3664. https://doi.org/10.3390/electronics12173664

Academic Editor: Donghyeon Cho

Received: 12 July 2023
Revised: 20 August 2023
Accepted: 25 August 2023
Published: 30 August 2023

1. Introduction

With the advancement of drone technology, drones are widely employed in various sectors, such as aerial photography, emergency response, and agricultural planning. However, the development of drones has also brought to the fore a series of management issues. These include illegal "rogue flights", the exploitation of drones for criminal and terrorist activities, and their potential to be transformed into dangerous weapons by carrying explosive materials [1–3]. Drones have become a new tool for terrorism, posing significant threats to public safety. In response to the increasingly severe UAV threat, it is urgent to establish an anti-drone system around restricted areas; thus, illegal UAV detection, as a critical component of the anti-drone system [4], has become a subject of widespread attention among researchers. Improving the accuracy and processing speed of detecting enemy UAV targets, conducting effective early warning detection, and then taking measures to intercept them is the key to mastering air control and maintaining national security and social stability. Most of the current early warning detection equipment has the defects of fixed deployment location, large size, and apparent target exposure, meaning that they cannot be flexibly distributed in hidden forward positions; therefore, lightweight and easy-to-deploy large-scale early warning equipment is needed to fill the gap. The following problems exist in solving the detection of UAV targets: (1) UAVs are characterized by their small size, the use of "stealth" materials, low-altitude reconnaissance targets, and flexible take-off platforms; (2) complex airspace environments are often affected by clouds, light, and object occlusion, so that the use of electromagnetic and other signals to detect UAV groups are prone to false detection and missed detection [5,6]. With the rapid development of computer vision technology and neural networks, methods based on video and image frames have been widely used to extract features such as target contours, colors, and shapes, enabling the real-time detection of target positions and motion behaviors. This approach has extensive applications in public security monitoring, intelligent transportation systems, national defense and security, human-computer interaction systems, and safety production. Applying computer vision technology to drone detection opens up a new avenue for airspace early warnings, offering vast prospects for practical applications [7].

The rest of the paper is structured as follows: Section 2 summarizes the works related to UAV detection. Section 3 first introduces the YOLOv8 network structure and the details of its critical modules, followed by an improved tiny UAV target detection model, and details the structure and roles of each improved module of the model. Section 4 first introduces the dataset and the experimental environment and then conducts ablation experiments, comparison experiments on the publicly available dataset TIB-Net, and, finally, self-built dataset experiments to validate the proposed method's feasibility fully. Section 5 summarizes the research results in the full paper and provides an outlook on future research directions.

2. Related Work

In recent years, improving hardware device performance has enhanced computer data-processing capabilities, enabling rapid advancements in visual technologies that rely on deep learning with big data. Object detection based on computer vision technology has garnered significant attention from researchers. It has evolved from traditional manual feature extraction [8–10] using convolutional calculations for object detection to leveraging deep learning to improve recognition accuracy in visual object detection. Compared to traditional electromagnetic signal detection methods such as radar, laser, infrared, audio, and radio frequency, object detection using visual sensors, specifically cameras capturing group videos and image data, offers more intuitive detection and the recognition of groups' information. It offers advantages such as the real-time and dynamic recording of sequential images of targets, low cost, fast detection speed, and immunity to interference from low-altitude clutter [11].

Object detection is an important research area in computer vision and is the foundation for numerous complex visual tasks. It has been widely applied in industries, agriculture,

and other fields [12,13]. Since 2014, there has been a remarkable advancement in deep learning-based object detection techniques. The industry has introduced various algorithms, including Faster R-CNN [14], SSD [15], and the YOLO series [16], to improve object detection further. With the rapid development of target detection technology, several useful methods have explicitly emerged for UAV target detection tasks [17–21]. For example, the authors of [17] argue that convolutional neural networks struggle to balance detection accuracy and model size. To address this issue, they introduced a recurrent pathway and spatial attention module into the original extremely tiny face detector (EXTD), enhancing its ability to extract features from small UAV targets. The model size is only 690.7 kb. However, this model exhibits a slow inference time and is unsuitable for deployment in practical engineering scenarios. Ref. [18] proposed a UAV target detection network based on multiscale feature fusion, which first extracts the target multisensory field features using res2net, then improves the network performance in terms of both fine-grained multiscale feature extraction and hierarchical multiscale feature fusion, and finally achieves better results on a self-built UAV detection dataset. Ref. [19] created a new UAV detection method that overcomes the limitations of the UAV detection process in terms of parameters and computational environment to perform realistic detection using web applications. In the current paper, we first screen an SSD pre-trained model that is suitable for deployment in this web application to improve detection accuracy and recall. The experimental results prove that the web application method outperforms the on-board processing method and achieves better results. Ref. [20] proposes a lightweight feature-enhanced convolutional neural network that is capable of the real-time and high-precision detection of low-flying objects. It effectively alerts against unauthorized drones in the airspace and provides guidance information. Ref. [21] introduces a novel deep learning method called the convolutional transformation network (CT-Net). The backbone of this network first incorporates an attention-enhanced transformation block, which establishes a feature-enhanced multi-head self-attention mechanism to improve the model's feature extraction capability. Then, a lightweight bottleneck module is employed to control computational load and reduce parameters. Finally, a direction feature fusion structure is proposed to enhance detection accuracy when dealing with multi-scale objects, especially small-sized objects. The approach achieves a mAP of 0.966 on a self-built low-altitude small-object dataset, demonstrating good detection accuracy. However, the FPS is only 37, indicating that there is room for improvement in detection speed.

Although significant progress has been made in UAV detection technology, existing detection methods still face challenges in balancing detection accuracy, model size, and detection speed. The YOLO series detection network has solved these problems effectively. The YOLO series models have undergone eight official iterations and several branch versions, showcasing remarkable detection accuracy and speed performance. These models have extensive applications in various fields, including medicine, transportation, remote sensing, and industry [22]. Scholars have extensively researched using the YOLO series models for UAV target detection, as evidenced by numerous studies [23–27]. For example, in reference [23], by incorporating an attention mechanism module into the PP-YOLO detection algorithm, enhancements were made to improve its performance. Furthermore, introducing the Mish activation function addressed the issue of gradient-vanishing during the backpropagation process, resulting in a significant boost in detection accuracy. In Ref. [24], a UAV detection algorithm for complex urban backgrounds was proposed, based on YOLOv3. It employed an FPN for multi-scale prediction, enhancing the system's detection performance for small targets. A lightweight Ghost network was also utilized to accelerate the model, achieving network lightweight status. Experimental results demonstrated that the algorithm effectively detected small UAV targets in complex scenes and exhibited strong robustness. In Ref. [25], a lightweight convolutional neural network, MobileNetv2, replaced the original CSPDarknet53 backbone of the high-performance YOLOv4 model. This substitution aimed to reduce the model's scale and simplify the computational operations. Experimental results demonstrated that Mob-YOLO could achieve accurate

real-time monitoring of UAV targets with smaller model sizes, making it deployable with onboard embedded processors. In Ref. [26], a YOLOv5-based distributed anti-drone system was proposed. This system integrates airport defense capabilities to address UAV jamming scenarios by incorporating features such as automatic targeting and jamming signal broadcasting, enabling the interception of illegal UAVs. To cater to the wide no-fly zone of the airport, the system is deployed around the airport using distributed clustering, effectively resolving the issues of blind detection and target loss. Experimental results have demonstrated the high accuracy of automatic targeting based on the YOLOv5 algorithm, with the inference speed and model size meeting real-time hardware detection requirements. Although the system needs to be more innovative to improve YOLOv5, the successful application of UAV target detection technology to practical engineering scenarios is also informative. Ref. [27] proposed the YOLOX-drone, an improved target detection algorithm for UAS based on YOLOX-S. Based on the YOLOX-S target detection network, this paper first introduces a coordinated attention mechanism to improve the image highlighting of UAV targets, enhance useful features, and suppress useless features. Secondly, for this paper, a feature aggregation structure has been designed to improve the representation of useful features, suppress interference, and improve detection accuracy. The improved algorithm performs well on both the publicly available DUT-Anti-AV dataset and the self-generated dataset, demonstrating its strong obstacle-detection capability.

Combining the improvement ideas proposed in the above-related literature on the YOLO series, this paper improves on the YOLOv8s model and offers a new model suitable for tiny UAV object detection, which achieves high detection accuracy and speed on the challenging small UAV dataset, and dramatically reduces the size of the model and the number of parameters. This study provides a new approach for model deployment in the field of tiny UAV object detection.

3. Methods

3.1. YOLOv8 Network Structure

YOLOv8 builds upon the success of previous versions of YOLO and introduces new features and improvements to enhance performance and flexibility further, achieving top performance and exceptional speed. YOLOv8 offers five different-sized models: nano, small, middle, large, and extra-large. The Nano model has a parameter count of only 3.2 million, providing convenience for deployment on mobile and CPU-only devices. In order to balance detection accuracy and speed, this paper employs YOLOv8s as the model for UAV detection, which is obtained by deepening and widening the nano network structure. YOLOv8 is divided into the backbone, neck, and head, which are used for feature extraction, multi-feature fusion, and prediction output. The design of the YOLOv8 network is shown in Figure 1.

The feature extraction network mainly extracts individual scale features from images created by the C2f and SPPF modules. The C2f module reduces the network by one convolutional layer based on the original C3 module, making the model more lightweight. It also incorporates the strengths of the ELAN structure from YOLOv7, effectively expanding the gradient branch using bottleneck modules to obtain richer gradient flow information [28]. SPPF reduces the network layers based on SPP (spatial pyramid pooling) [29] to eliminate redundant operations and perform feature fusion more rapidly. The multiscale fusion module adopts a combination of an FPN (feature pyramid network) [30] and PAN (path aggregation network) [31]. By bi-directionally fusing the low-level features and high-level features, it enhances low-level features with smaller receptive fields and improves the detection capability of targets at different scales. The detection layer predicts target positions, categories, confidence scores, and other information. The head part of YOLOv8 switches from an anchor-based to an anchor-free approach. It abandons the IOU matching or single-side scale assignment and uses the task-aligned assigner for positive and negative sample matching. Ultimately, it performs multi-scale predictions using 8×, 16×, and

32× down-sampled features to achieve accurate predictions for small, medium, and large targets. The detailed modules in the YOLOv8 network are illustrated in Figure 2.

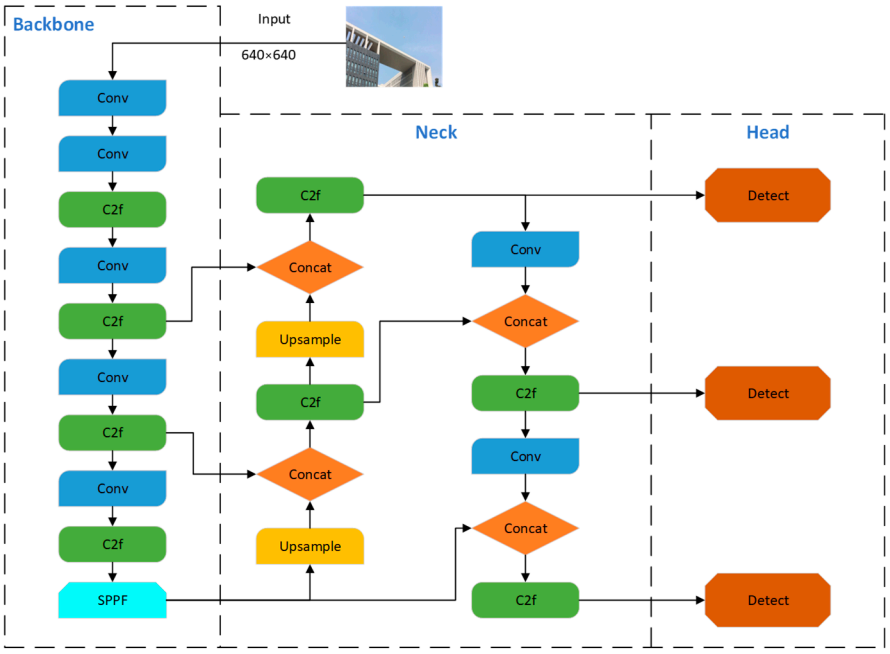

Figure 1. YOLOv8 network structure diagram.

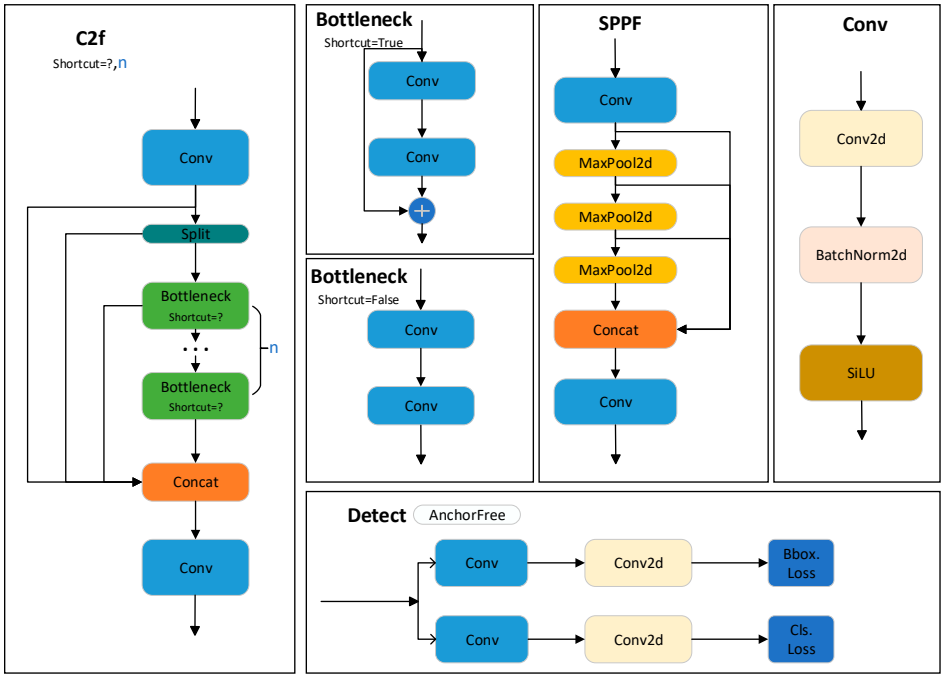

Figure 2. YOLOv8 network detail structure diagram.

3.2. Improved YOLOv8 UAV Detection Model

YOLOv8 extracts the target features by using a deep residual network. It completes the multiscale prediction using the PAN structure, but YOLOv8 still performs three down-sampling iterations when extracting features to obtain the maximum feature map. However, much of the target feature information is lost, which could be useful for detecting tiny targets. Therefore, this paper improves YOLOv8 and proposes a network model for UAV micro-target detection, and the improved network structure is shown in Figure 3. The

specific improvement schemes are as follows. (1) We enhanced the detection capability of the model for tiny targets by adding a high-resolution detection branch in the detection head part; meanwhile, the detection layer and its related feature extraction and fusion layer for large target prediction were cut, and the model parameters were reduced. (2) The multiscale feature extraction module was improved by using SPD-Conv [32] instead of Conv to extract multiscale features. (3) The GAM attention mechanism [33] was introduced into the multiscale fusion module to enhance the model's fusion of target features.

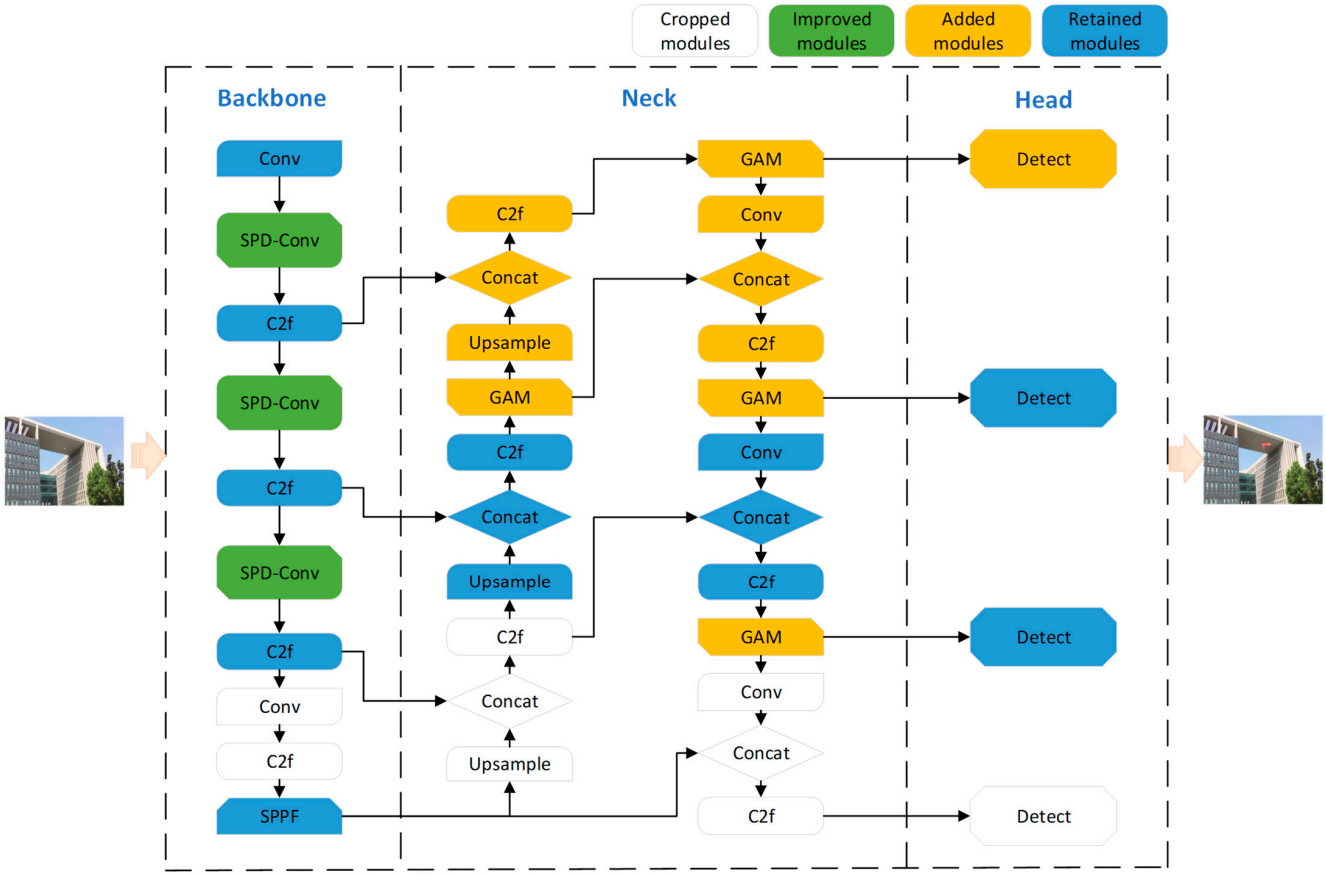

Figure 3. Improved YOLOv8 network structure diagram.

3.2.1. Improvement of the Detection Head

A. Adding a tiny-target detection head

In this paper, the detection object is a low-flying UAV. When using the camera to capture the UAV image, in order to prevent the flying UAV from rushing out of the camera's field of view, the camera generally maintains a large area of view. Hence, the proportion of the UAV in the image is usually small. The original YOLOv8 model backbone network down-samples for a total of five times to obtain five layers of feature expressions (P1, P2, P3, P4, and P5), wherein Pi denotes a resolution of 1/2i of the original image. Although multi-scale feature fusion is achieved in the neck network via top-down and bottom-up aggregation paths, this does not affect the scale of the feature map, and the final detection head part is detected after passing through P3, P4, and P5. The feature map scales are 80 × 80, 40 × 40, and 20 × 20, respectively. In the small target detection task, there are often tiny targets to be detected. The TIB-Net data used in this paper contains many tiny UAV targets, usually smaller than 10 × 10 pixels in scale. Such marks have lost most of their feature information after multiple down-sampling and are still challenging to detect with high resolution by the P3 layer detection head.

To achieve micro-target identification, as mentioned above, and also gain a better detection effect, we introduced a new detection head on the YOLOv8 model by P2 layer features, called the micro-target detection head; the structure is shown in Figure 4. The resolution of the P2 layer detection head is 160 × 160 pixels, which is equivalent to only two down-sampling operations in the backbone network, containing richer information on the underlying features of the target. The two P2 layer features, obtained from top-down and bottom-up in the neck network, are fused with the same scale features in the backbone network, in the form of concat, while the output features are the fused results of the three input features, which makes the P2 layer detection head fast and effective when dealing with tiny targets. The P2 layer detection head, together with the original detection head, can effectively mitigate the scale variance caused by the P2 detection head, which, together with the initial detection head, can effectively reduce the negative effects of scale variance. The added detection head is specific to the underlying features and is generated from low-level, high-resolution feature maps, which are more sensitive to small targets. Although adding this detection head increases the computation and memory overhead of the model, it significantly improves the detection of tiny targets.

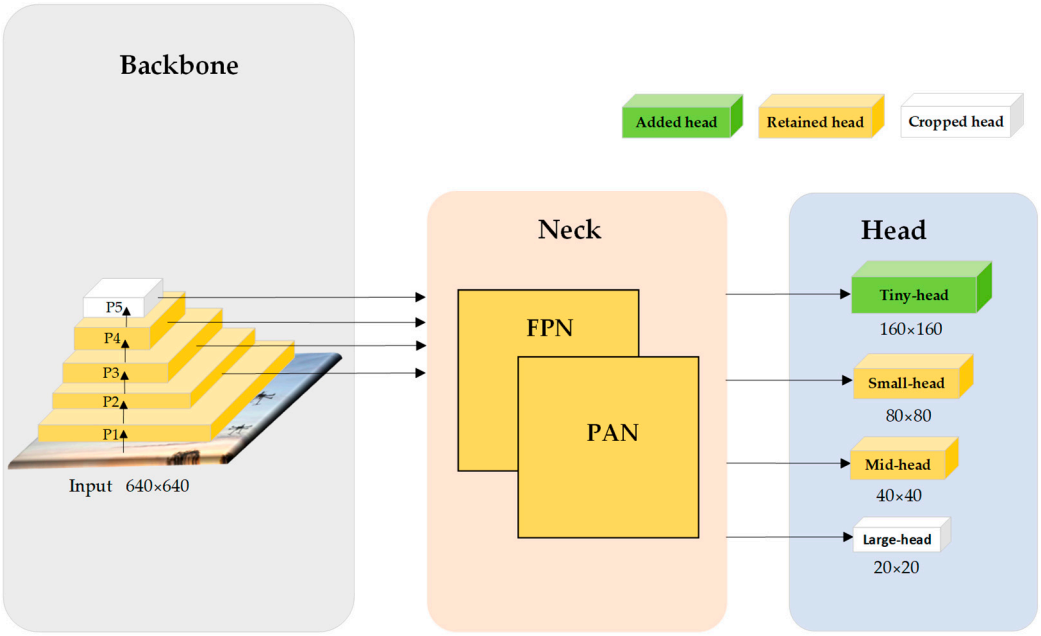

Figure 4. Improvement scheme at the head.

B. Removing the large-target detection head

The large target detection header P5 layer is obtained by down-sampling the image by a factor of 32. When the target size is smaller than 32 pixels, it is likely that, at most, only one point of the target is sampled or not sampled. Therefore, the YOLOv8 large target detection layer is redundant when detecting small-sized UAV targets. Based on the above conclusions, this paper cuts out the large target prediction layer and the related feature extraction and feature fusion layers from the YOLOv8 network structure. It only retains the 4-fold down-sampling, 8-fold down-sampling, and 16-fold down-sampling feature maps for UAV prediction. In the improved network structure shown in Figure 3, the 16-fold down-sampled feature maps of the third C2f layer are directly fed into SPPF for multi-scale feature extraction. The fused feature maps are then discarded from the Upsample-Concat-C2f module and directly connected to the next module, and all network layers after the medium target detection layer are discarded. This improved network structure reduces the computational bottleneck by removing redundant calculations with guaranteed accuracy. The improved detection head is shown in Figure 4.

3.2.2. Improvement of the Feature Extraction Module

When the image shows good resolution, and the detection object is of moderate size, the image contains a significant enough amount of redundant pixel information that strode convolution (i.e., stride > 1) can conveniently skip this redundant pixel information. The model is still able to learn features efficiently. However, in more complex tasks involving ambiguous images and small objects, the assumption of redundant information no longer holds, and the current model starts to suffer from a loss of detail, which significantly impairs its ability to learn features. Small objects are challenging to detect because they are characterized by low resolution and have limited information about the content needed to learn patterns. In YOLOv8, the feature extraction module Conv, a stride convolutional layer, rapidly degrades its detection performance in tasks with low image resolution or small detection objects. For this reason, the current paper introduces a new CNN building block, SPD-Conv, in the feature extraction stage to replace the stride convolution layer. SPD-Conv consists of an SPD (space-to-depth) layer and a non-stride convolution layer and can be applied to most CNN architectures. In an earlier study [32], the authors introduced SPD-Conv into the backbone and neck of YOLOv5. They experimentally demonstrated that the method significantly improved the performance in complex tasks dealing with low-resolution images and small objects. Combined with the improved ideas of this paper for YOLOv5, demonstrated experimentally, we only need to introduce SPD-Conv in the feature extraction module (i.e., backbone) of YOLOv8 to improve the detection of tiny UAV targets without adding too much redundancy, as shown in Figure 3. The SPD-Conv structure is shown at a scale = 2 in Figure 5.

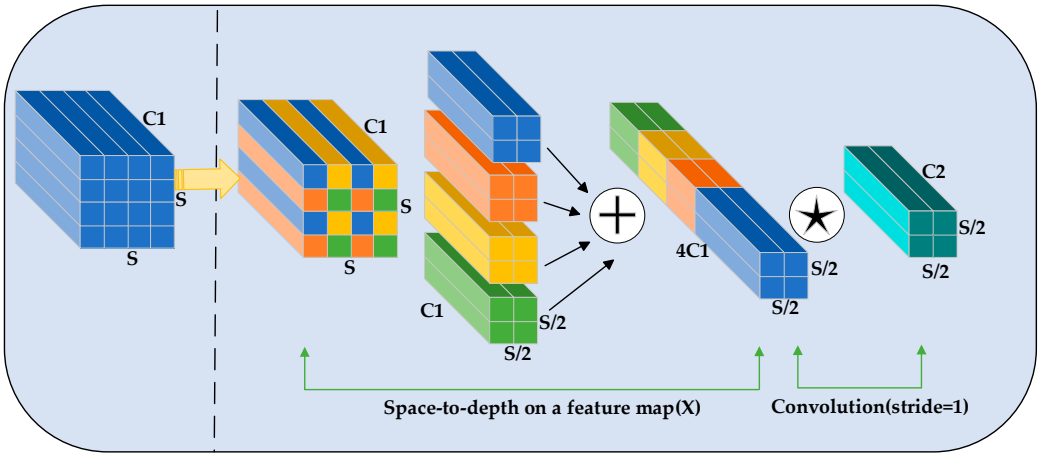

Figure 5. Structure of SPD-Conv.

The SPD-Conv operation consists of two steps. Firstly, the feature map of the input image undergoes preprocessing from space to depth; subsequently, the preprocessed feature map is subjected to a standard convolution. Figure 5 illustrates the feature map of a C1 channel, demonstrating the process of slicing up the input feature map. After pruning, four sets of sub-shaped images are obtained, where each sub-shaped image retains the same number of channels as the input feature map. As the scale is set to 2, the width and height of the output feature map are halved compared to the input. The resulting sub-feature images are combined through a standard convolution, ensuring the preservation of all sub-feature information due to the use of a standard convolution with a step size of one.

3.2.3. Improvement of the Feature Fusion Module

GAM, an attention mechanism module, is a lightweight, practical, and simple component that can be seamlessly integrated into CNN architectures. Its primary purpose is to enhance the performance of deep neural networks by minimizing information loss and amplifying global interaction representation within a given feature mapping. The

GAM module adopts the CBAM attention mechanism, which operates from channel to spatial order. In an earlier work [33], the GAM module was successfully integrated into various models across different datasets and classification tasks, resulting in significant improvements in model performance that underscore the efficacy of the GAM module. As a plug-and-play module, GAM is widely cited, as in the literature [34], by inserting GAM into the backbone and head of YOLOv7, enabling the network to extract critical features by amplifying the interaction of global dimensional features. The GAM structure is shown in Figure 6.

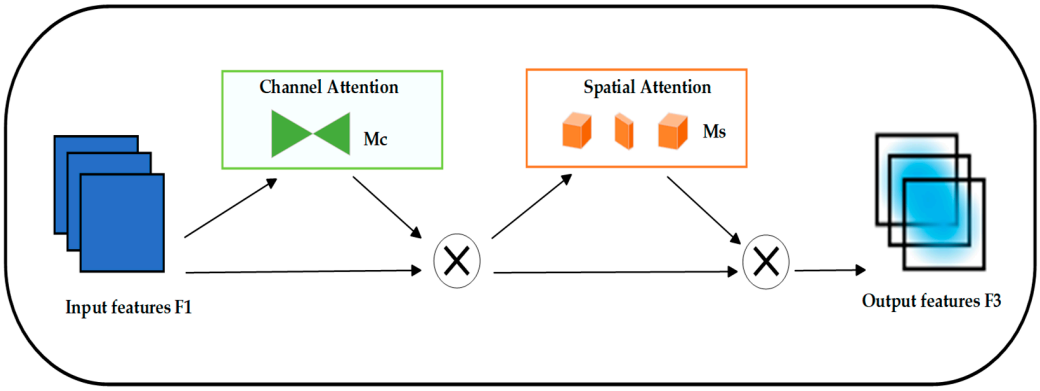

Figure 6. The GAM attention module.

Given the mapping of input attribute F_1, intermediate states F_2 and output F_3 are defined as follows:

$$F_2 = M_c(F_1) * F_1 \quad (1)$$

$$F_3 = M_s(F_2) * F_2 \quad (2)$$

Since small targets are small in size and have few and inconspicuous features, adding the GAM attention module to the feature fusion network can amplify global interaction and enhance the retention ability of the network for small target features, while directly improving the feature fusion in the neck part of the network. In the detection task, the GAM attention module can help the model to extract the attention region effectively and improve the detection performance.

4. Experimental Preparation and Results

In this paper, we use the public UAV dataset TIB-Net [17] to evaluate the model's performance and introduce the dataset, network setup and training, evaluation index, ablation experiment, comparison experiment, and self-built dataset experiment.

4.1. Dataset Introduction

The TIB-Net UAV dataset comprises 2850 images showcasing various types of UAVs, including multi-rotor UAVs and fixed-wing UAVs. The images were captured by a fixed camera on the ground at a distance of about 500 m from the aerial drones, and the resolution of the collected images was 1920 × 1080 pixels. These scenes cover several low-altitude scenes (sky, trees, buildings, etc.) from UAV flight images, fully considering samples at different times of the day and in different weather. It can be seen from Figure 7 that the UAV occupies only less than 1% of each image. Some of the samples are shown in Figure 8.

4.2. Network Setup and Training

This section details the training process of the TIB-Net dataset on YOLOv8 and the modified YOLOv8. The hardware configuration used for the experiments is an 8 GB NVIDIA GeForce RTX 3070 graphics card, the deep learning framework PyTorch 1.13.1, Python version 3.7.15, CUDA version 11.7, and Ubuntu 22.04 as the operating system.

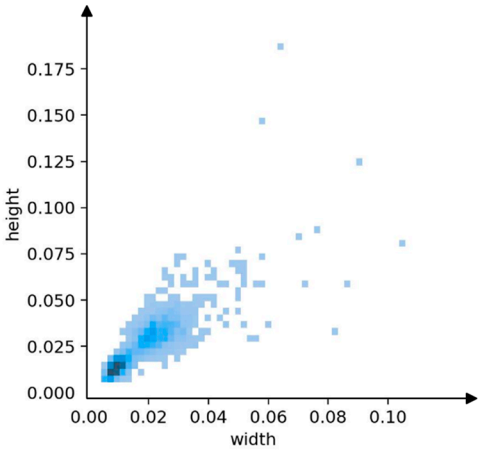

Figure 7. Proportion of drone size in the image (darker colors mean more drones).

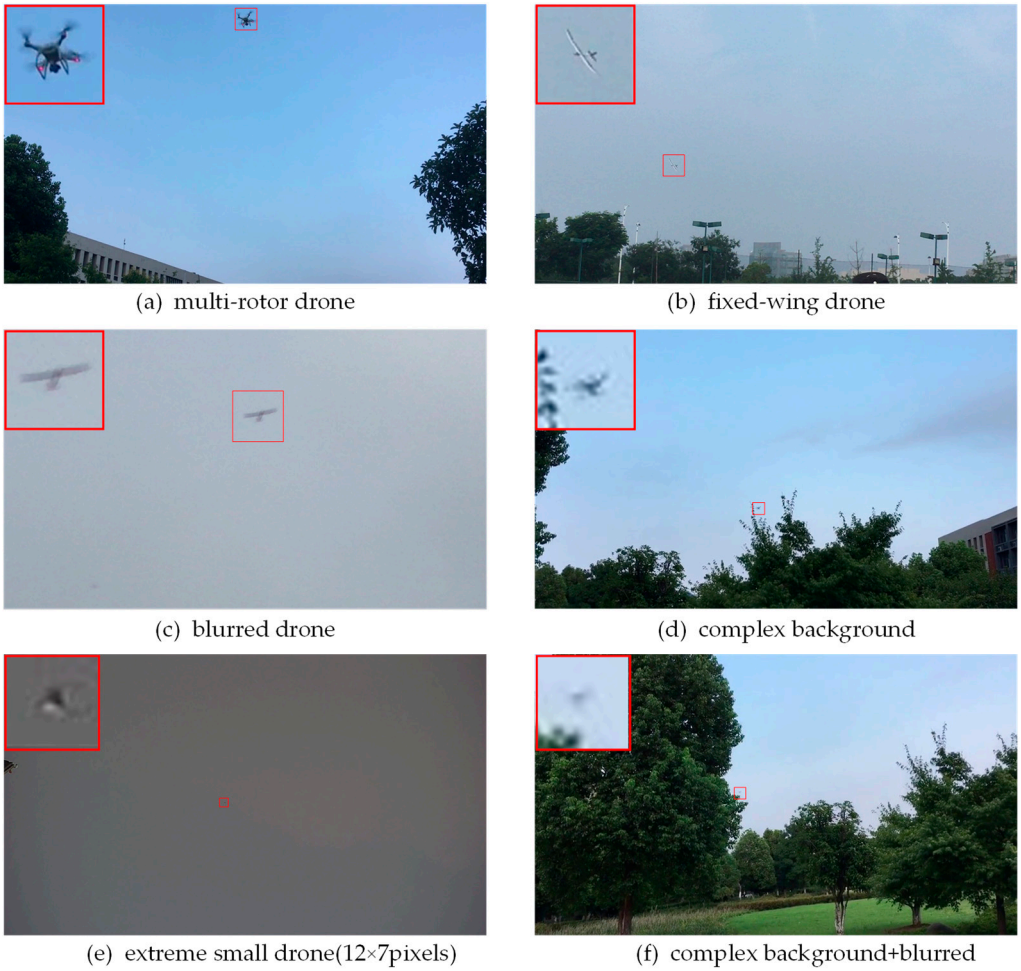

Figure 8. Display of dataset diversity. (**a**) multi-rotor drone; (**b**) fixed-wing drone; (**c**–**f**) show several difficult samples, which contain extreme small drone, blurred drone or complex environment.

4.2.1. Loss Function Setting

The loss functions of the improved YOLOv8 are consistent with YOLOv8, and both include rectangular box loss ($Loss_{box}$), distribution focus loss ($Loss_{dfl}$), and classification loss ($Loss_{cls}$).

$$Loss = a \cdot Loss_{box} + b \cdot Loss_{dfl} + c \cdot Loss_{cls} \tag{3}$$

Among them, a, b, and c all represent the weight proportion of the corresponding loss function in the total loss function. In this experiment, the three weights are a = 7.5, b = 1.5, and c = 0.5, respectively.

4.2.2. Network Training

Before training, the dataset images and labels are divided into the training set, validation set, and test set in a ratio of 7:1:2. The maximum number of epochs for training the dataset is set to 150, with the first three epochs used for warm-up training. The SGD optimization strategy is employed for learning rate adjustment, with an initial learning rate of 0.01. Considering the presence of numerous tiny objects in the sample images and the need to balance real-time performance with accuracy in the detection process, the sample size is normalized to 640 × 640. This size allows the model to be deployed on edge devices without losing too much helpful information from the images. To ensure fairness and the comparability of the model's performance, no pre-trained weights are used in ablation or comparative experiments. Additionally, all training processes share consistent hyperparameter settings. The most important parameter settings for the training process are shown in Table 1.

Table 1. Important parameter setting table.

Parameters	Setup
Epochs	150
Warmup-epochs	3
Warmup-momentum	0.8
Batch Size	8
Imgsize	640
Initial Learning Rate	0.01
Final Learning Rate	0.01
Patience	50
Optimizer	SGD
NMSIoU	0.7
Momentum	0.937
Mask-ratio	4
Weight-Decay	0.0005

4.3. Evaluation Indicators

To validate the model performance, P, R, AP, mAP, the number of parameters, model size, and frames per second (FPS) [35] are chosen as experimental evaluation indicators.

(1) Accuracy and recall rates are calculated as follows:

$$P = \frac{TP}{TP + FP} \cdot 100\% \tag{4}$$

$$R = \frac{TP}{TP + FN} \cdot 100\% \tag{5}$$

where TP (true positives) denotes the number of targets detected correctly, FP (false positives) denotes the number of backgrounds detected as targets, and FN (false negatives) denotes the number of targets detected as backgrounds.

(2) The average precision and average precision mean are calculated as follows:

$$AP = \int_0^1 p(r)d(r) \tag{6}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

where N is the number of categories and AP is the average accuracy of each category. In our UAV detection task, N = 1.

4.4. Ablation Experiments

For this section, based on the TIB-Net UAV dataset, ablation experiments were conducted to explore the improvement effects of each added or modified module on the overall model. Starting with the original YOLOv8s as a baseline, the detection head, backbone, and neck improvements were sequenced. To analyze the performance improvement of each module, the benchmark Model 1, improved Model 2 (with added tiny-head), improved Model 3 (added tiny-head and cropped large-head), improved Model 4 (with added tiny-head, cropped large-head, and improved SPD-Conv), improved Model 5 (with added tiny-head, cropped large-head, and added GAM), and improved Model 6 (with added tiny-head, cropped large-head, improved SPD-Conv, and added GAM) were defined. The changes in evaluation metrics for these six models were quantitatively explored, and the optimal results for each evaluation metric were highlighted. The experimental results of the models on the TIB-Net dataset are shown in Table 2.

Table 2. Results of the various ablation experiments.

Components	**1**	**2**	**3**	**4**	**5**	**6**
+Tiny-Head		√	√	√	√	√
-Large-Head			√	√	√	√
+SPD-Conv				√		√
+GAM					√	√
P	81.4%	92.2%	91.9%	92.6%	93.1%	**93.3%**
R	78.1%	91.6%	91.6%	92.8%	92.2%	**93.3%**
mAP	86.1%	94.4%	93.5%	94.9%	93.6%	**95.1%**
Parameters/million	11.126	10.852	**3.527**	4.209	3.785	4.467
Model Size/MB	21.9	22.1	**7.3**	8.7	7.9	9.2
FPS/f.s-1	**285**	217	259	232	246	221

Referring to Table 2, it can be seen that:

1. The increase from the tiny detection head improved the model by 10.8%, 13.5%, and 8.3% for P, R, and mAP, respectively, indicating that the increase from the high-resolution detection head can effectively enhance the detection ability of tiny targets. At the same time, it can be seen that after trimming off the large target detection layer, the parameter amount was reduced by 70.2% and the model size was reduced by 67%, while R remained unchanged, P was reduced by 0.3%, and mAP was reduced by 0.9%, indicating that the low-resolution detection head made little contribution to the detection of tiny UAV targets and generated a large redundant network.
2. The experimental results of improving models 3, 4, 5, and 6 show that improving the SPD-Conv module had a better improvement effect on the recall R of the model, indicating that improving the Conv module to SPD-Conv in the backbone network can better retain the features of the minutiae targets and reduce the probability of missing detection for the minutiae targets; adding GAM had a better improvement effect on the accuracy P of the model, indicating that adding the GAM attention module in the addition of the GAM attention module in the neck had a good impact on the feature fusion of the network and reduced the probability of false network detection. When

both SPD-Conv and GAM were added, P, R, and mAP were improved, although the number of parameters and the model size slightly increased.

3. Comparing the experimental results of the improved model 6 (i.e., our model) and model 1 (i.e., the base model), as shown in Figure 9, we can see that because the tiny-head, SPD-Conv, and GAM modules added some inference time, the improved model FPS metric reached 221/f.s-1, which is lower compared to the 285/f.s-1 of the base model; however, it can still guarantee meeting the real-time requirement in actual deployment. In addition, our model significantly improved the P, R, mAP, number of parameters, and model size compared with the base model, with P, R, and mAP improving by 11.9%, 15.2%, and 9%, respectively. The number of parameters and model size decreased by 59.9% and 57.9%, respectively, thus proving the effectiveness and practicality of the improved model.

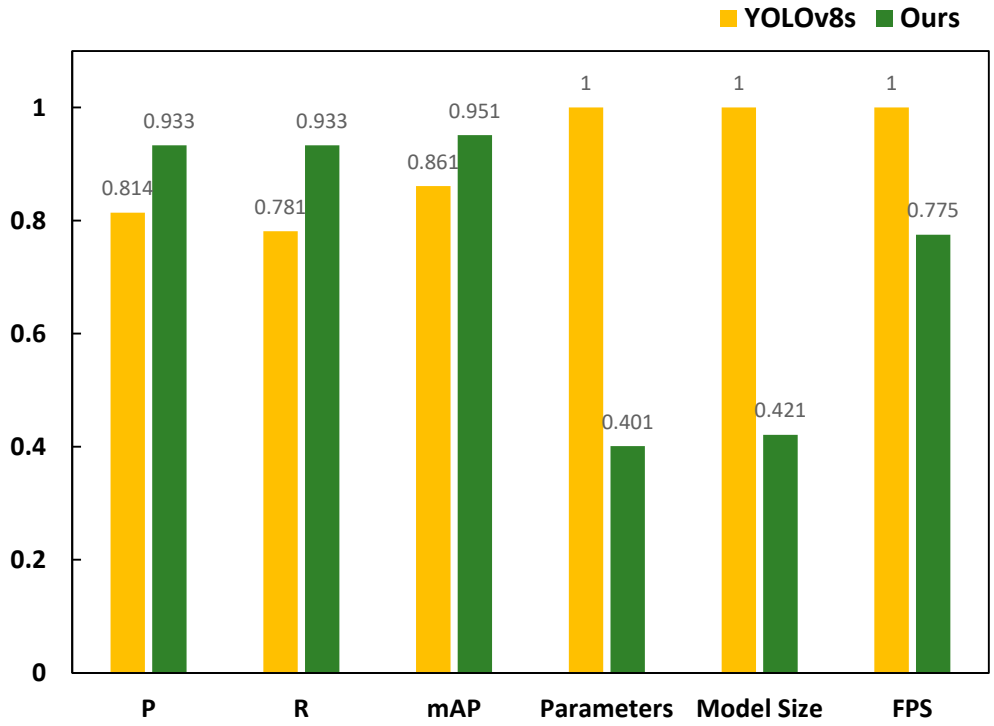

Figure 9. Comparison graph between our model and the YOLOv8s experiment (parameters, model size, and FPS are normalized separately).

In order to observe the detection effect of the improved model more intuitively, the base model YOLOv8s and the improved model in this paper are used for drone detection, and the effect comparison graphs are shown in Figures 10 and 11, respectively. In Figures 10 and 11, the detection results of YOLOv8s are shown on the left, and the detection results of the improved model are shown on the right. The UAV position and confidence level are indicated by rectangular boxes and text, respectively, and the details of the area where the UAV is located are shown in the upper right corner or lower right corner of the images, respectively.

In Figures 10 and 11, a comparison reveals that YOLOv8s exhibit instances of missed detections when the UAVs are very small or have blended into the background, as shown in Figure 10a,c,e, while false detections as shown in Figure 11a,c,e, highlighted by the yellow boxes. In contrast, the improved model proposed in this paper accurately detects small UAV targets against complex backgrounds such as buildings and trees. Additionally, our method significantly improves the confidence regarding the detected UAVs. As shown in Figure 10b, the confidence reached 0.96, while, as shown in Figure 11e,f, the confidence increased from 0.27 to 0.82. Therefore, the improved model in this paper effectively addresses the issues of missed and false detections of small UAV targets against complex backgrounds.

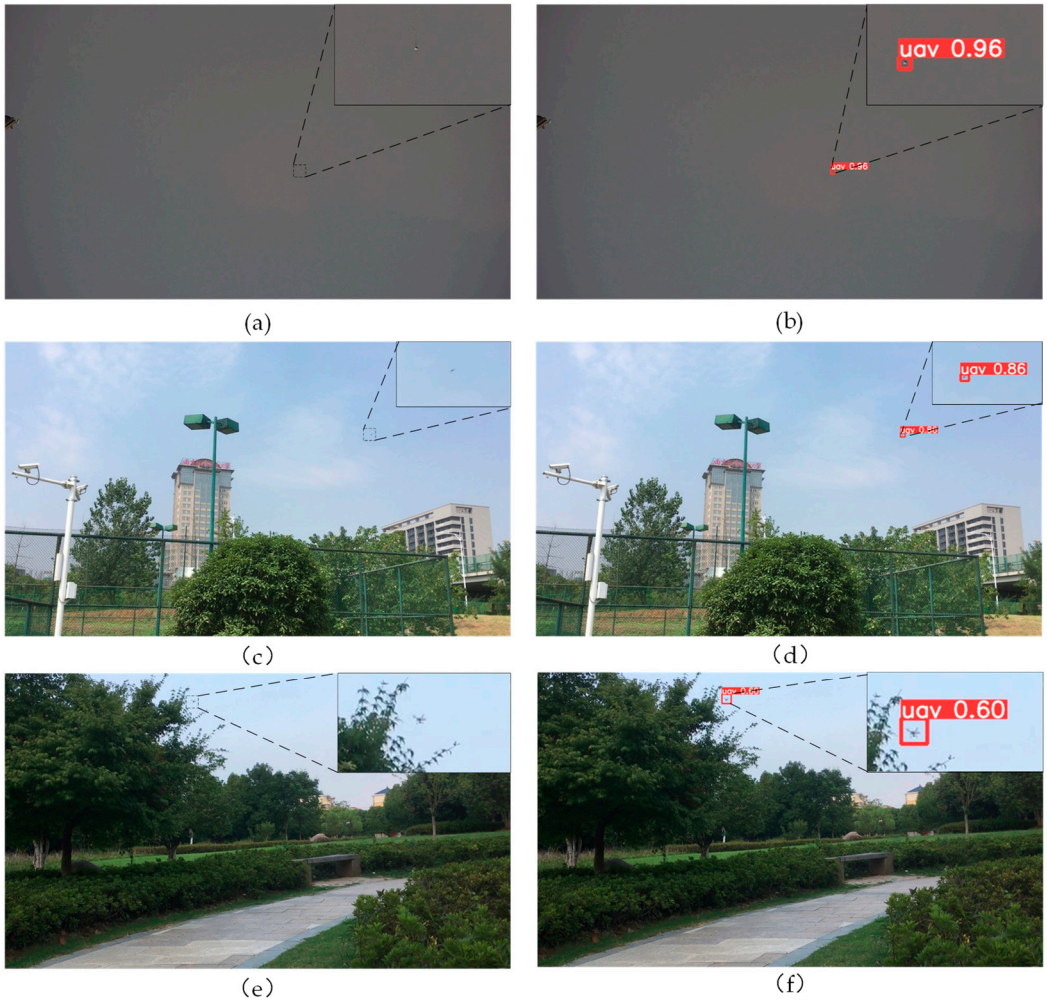

Figure 10. The left side shows some of the leakage detection results of YOLOv8s, as shown in Figure (**a**,**c**,**e**). The right side shows the detection results of the improved model in the same image, as shown in Figure (**b**,**d**,**f**).

4.5. Comparative Experiments

To further verify the advantages of the algorithm used in this paper, the algorithm in this paper was compared with other YOLO series algorithms for experiments, and four advanced YOLO series algorithms (YOLOv5-S [36], YOLOX-S [37], YOLOv7 [38], YOLOv7-tiny) at the present stage were selected on the TIB-Net dataset, taking into account the lightweight model size and detection performance, respectively. To fully reflect the model's superiority in this paper, the TIB-Net [17] model was also selected as a comparison object in the experiments. The parameters of the comparison experiments were carried out according to Table 1, and the evaluation metrics were consistent with Table 3. The selected experimental models are all official versions. The results of the comparison experiments are shown in Table 3.

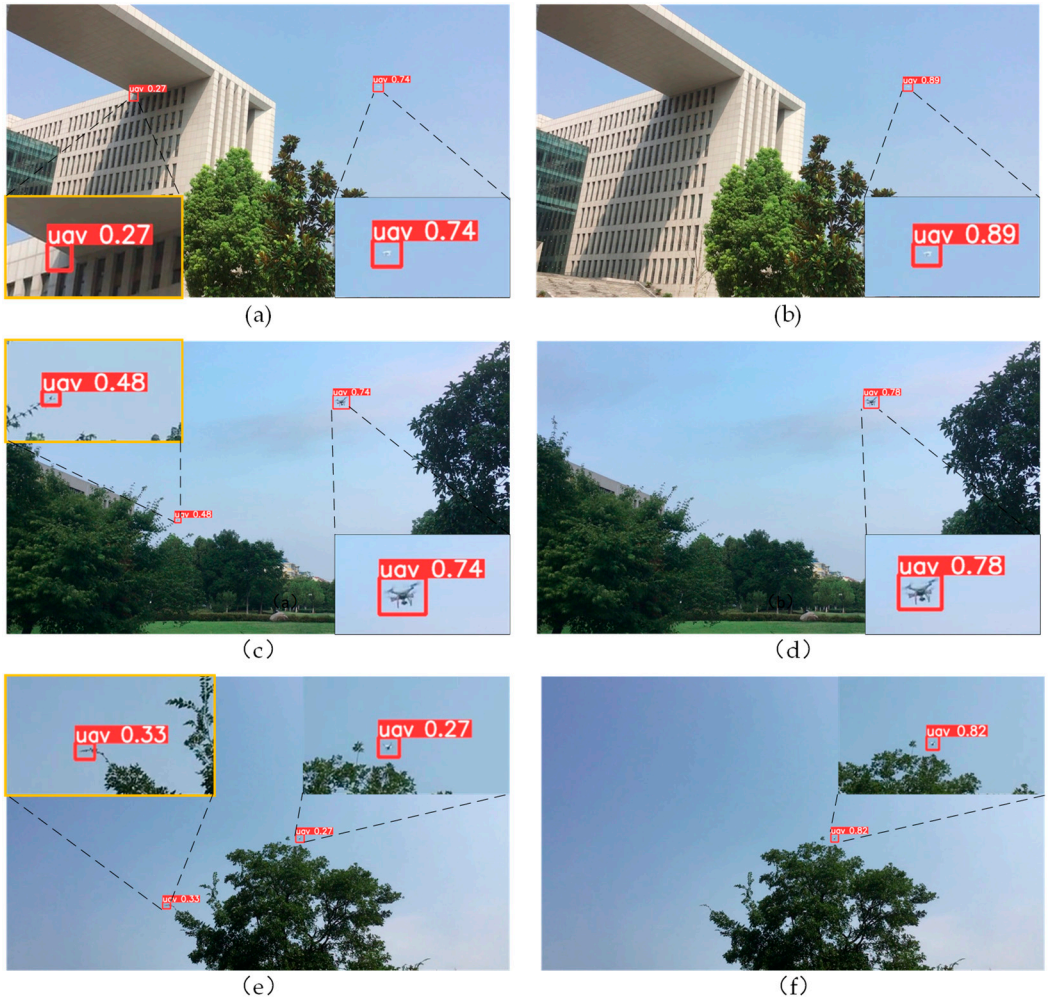

Figure 11. Figure (**a**,**c**,**e**) show the results of the partial error detection of YOLOv8s, as shown in the yellow box, and Figure (**b**,**d**,**f**). show the detection results of the improved model for the same image.

Table 3. Comparison of experimental results.

Methods	P	R	mAP	Parameters/Million	Model Size/MB	FPS/f.s-1
TIB-Net	87.6%	87%	89.4%	**0.163**	**0.681**	5
YOLOv5-s	88.1%	90.9%	91.2%	7.013	14.3	**256**
YOLOX-s	90.5%	80.6%	88.7%	9.0	62.5	132
YOLOv7	64.2%	56%	52.4%	36.480	74.7	104
YOLOv7-tiny	85%	82.6%	85%	6.007	12.2	227
Ours	**93.3%**	**93.3%**	**95.1%**	4.467	9.2	221

According to Table 3, it can be seen that:

1. Comparing YOLOv7 and YOLOv7-tiny, it can be seen that although the number of parameters and the model size of YOLOv7 are much higher than the other models, P, R, and mAP present the worst results. Conversely, YOLOv7-tiny achieves good results in terms of detection accuracy, with a smaller number of parameters and model size. The reason for this is that the TIB-Net dataset has a smaller drone size and has fewer drone features contained in the images, while the more complex YOLOv7 network structure may learn many useless background features, which, in turn, results in poorer detection results.
2. The TIB-Net detection network is at the other extreme; it can still maintain better detection accuracy with a much smaller number of parameters and model size than

other models. However, one disadvantage is also apparent; the FPS is only 5, far from meeting the needs of real-time UAV detection.

3. YOLOv5-s yields the best overall performance except for our model, while the FPS is 256 ahead of all models, and the P and R values are well balanced. In addition, the detection of YOLOX is also good, but R and FPS are slightly low compared with YOLOv5-s, and the model size is too large.
4. The improved model proposed in this paper outperforms other models in terms of P, R, and mAP. In addition, it is at the top of all the models in terms of the number of parameters, model size, and FPS, while the number of parameters and model size is only higher than the TIB-Net network; FPS is slightly lower compared to YOLOv5-s and YOLOv7-tiny, but it can meet the deployment requirements of real-time detection. Overall, the tiny UAV detection network proposed in this paper achieves better detection accuracy, model size, and detection speed and can meet the specifications of practical engineering applications.

4.6. Self-Built Dataset Experiment

In order to evaluate the generalization performance of the model, this paper used cameras to collect UAV flight images on different scenes and different periods and collected a total of 1091 images of low-altitude scenes of various models of UAVs from major video sites such as YouTube and other web channels to make a new dataset. Figure 12 shows that most of the drones in the self-built dataset also occupy less than 1% of each image, compared with Figure 7, where this is larger than for the drones in the TIB-Net dataset. In addition, many new UAV images taken from high altitudes were added, to increase the diversity of the dataset. Compared with the TIB-Net dataset, where most of the dataset images are set against the sky, the background of the self-constructed dataset is more complex, as shown in Figure 13, where the drone blends in with the mountain or plants.

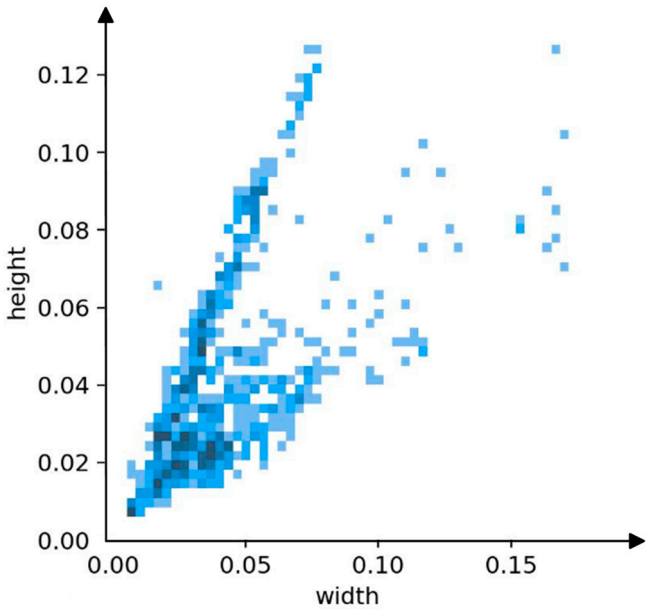

Figure 12. Size of self-built dataset drones (darker colors mean more drones).

In the self-built dataset experiments, the new dataset was divided into training and validation sets in the ratio of 7:3. To be consistent with the TIB-Net dataset, the images were first resized to 640 × 640 for training, and the training parameters were consistent with those in Table 1. The experimental results are shown in Table 4 and Figure 14.

Figure 13. Selected sample plots of the self-built dataset. (**a**–**c**) show drone imagery from different time periods; (**d**–**i**) show several difficult samples, including very small drones, drones photographed from a high altitude, or complex environments.

Table 4. Self-built dataset comparison—experimental results.

Model	P	R	mAP
YOLOv8s	88.8%	73.9%	85.2%
Ours	97%	89.5%	95.3%

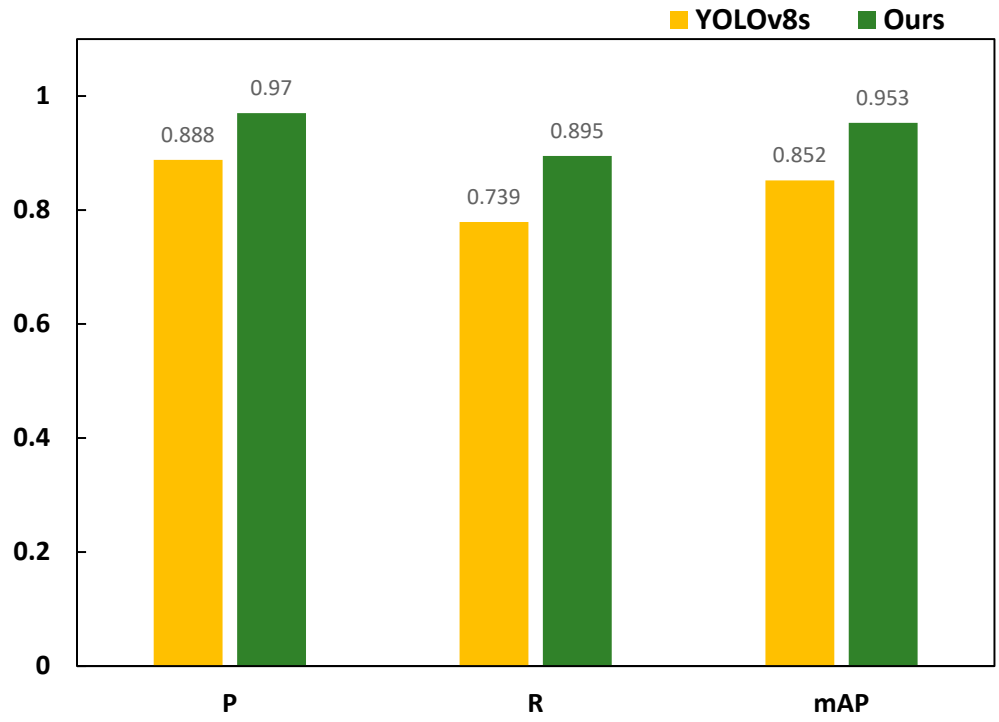

Figure 14. Comparison graph between our model and the YOLOv8s experiment (self-built datasets).

As can be seen from Table 4, the P, R, and mAP of the improved model with the new dataset were 97%, 89.5%, and 95.3%, respectively, which were about 8.2%, 15.6%, and 10.1% higher, respectively, compared to the pre-improvement period. Comparing Tables 2 and 4, it can be seen that the improved model improved P by 3.7% in the new dataset because the UAV target volume in the new dataset was generally larger than that in the TIB-Net dataset. However, the picture background in the new dataset was more complex. Hence, the improved model reduced R by 3.8% in the new dataset. Overall, the improved model still has high detection accuracy and shows that our method has good generalization. The actual detection results are shown in Figure 15.

Figure 15. Actual test chart display.

5. Conclusions and Outlook

To address the problem that tiny UAV targets are challenging to detect, this paper proposes an improved YOLOv8 detection model that can accurately detect UAV image targets while satisfying edge device deployment. The model overcomes the adverse effects of UAV size, airspace background, light intensity, and other factors on the detection task. Specifically, firstly, in the detection head part, the high-resolution detection head is added to improve the detection capability regarding tiny targets. In contrast, the large target detection head and redundant network layers are cut off to effectively reduce the number of network parameters and improve the UAV detection speed. Finally, the GAM

attention mechanism is introduced in the neck to improve the target feature fusion of the model, thus improving the model's overall performance for UAV detection. Ablation and comparison experiments were conducted on a complex TIB-Net dataset. Compared with the baseline model, our method improved P, R, and mAP by 11.9%, 15.2%, and 9%, respectively. Meanwhile, the number of parameters and model size were reduced by 59.9% and 57.9%, respectively. In addition, the detection model achieved better results in the comparison experiments and self-built dataset experiments. In conclusion, our method is more suitable for engineering deployment and the practical application of UAV target detection systems.

However, due to adding extra detection heads in the model and using both SPD-Conv and GAM modules, which increased the model inference time, the FPS decreased compared to the baseline model. In addition, from the self-built dataset experiments, it can be seen that R decreases when the airspace background is more complex, i.e., the probability of missing detection increases. Follow-up work will then be devoted to improving the detection accuracy in more complex airspace backgrounds while reducing the model inference time.

Author Contributions: Conceptualization, X.Z. and Z.H.; methodology, X.Z.; software, X.Z.; validation, X.Z., Z.H. and S.W.; formal analysis, T.L.; investigation, T.L.; resources, H.L.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z.; visualization, X.Z.; supervision, Z.H.; project administration, X.Z. and H.L.; funding acquisition, Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Xinjiang Uygur Autonomous Region of China, grant number 2022D01C59.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: More information is available at https://github.com/kyn0v/TIB-Net (accessed on 29 July 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shi, X.; Yang, C.; Xie, W.; Liang, C.; Shi, Z.; Chen, J. Anti-Drone System with Multiple Surveillance Technologies: Architecture, Implementation, and Challenges. *IEEE Commun. Mag.* **2018**, *56*, 68–74. [CrossRef]
2. Chen, Q.Q.; Feng, Z.W.; Zhang, G.B. Dynamic modelling and simulation of anti-UAV tethered-net capture system. *J. Natl. Univ. Def. Technol.* **2022**, *44*, 9–15.
3. Ikuesan, R.A.; Ganiyu, S.O.; Majigi, M.U.; Opaluwa, Y.D.; Venter, H.S. Practical Approach to Urban Crime Prevention in Developing Nations. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security, Marrakech, Morocco, 31 March–2 April 2020; pp. 1–8.
4. Mahmood, S.A. Anti-Drone System: Threats and Challenges. In Proceedings of the 2019 First International Conference of Computer and Applied Sciences (CAS), Baghdad, Iraq, 18–19 December 2019; p. 274.
5. Wu, X.; Sahoo, D.; Hoi, S.C.H. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]
6. Zhao, Z.Q.; Zheng, P.; Xu, S.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
7. Garcia, A.J.; Lee, J.M.; Kim, D.S. Anti-drone system: A visual-based drone detection using neural networks. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 21–23 October 2020; pp. 559–561.
8. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
9. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
10. Bay, H.; Tuytelaars, T.; van Gool, L. Surf: Speeded up robust features. In Proceedings of the Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Proceedings, Part I 9. Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
11. Dai, J.; Wu, L.; Wang, P. Overview of UAV Target Detection Algorithms Based on Deep Learning. In Proceedings of the 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 17–19 December 2021; Volume 2, pp. 736–745.

12. Zuo, Y. Target Detection System of Agricultural Economic Output Efficiency Based on Kruskal Algorithm. In Proceedings of the 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, India, 2–3 December 2022; pp. 1–5.
13. Li, S.; Yu, J.; Wang, H. Damages detection of aero-engine blades via deep learning algorithms. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5009111. [CrossRef]
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
16. Jiao, L.C.; Zhang, F.; Liu, F.; Yang, S.Y.; Li, L.L.; Feng, Z.X.; Qu, R. A Survey of Deep Learning-Based Object Detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
17. Sun, H.; Yang, J.; Shen, J.; Liang, D.; Ning-Zhong, L.; Zhou, H. TIB-Net: Drone Detection Network with Tiny Iterative Backbone. *IEEE Access* **2020**, *8*, 130697–130707. [CrossRef]
18. He, J.; Liu, M.; Yu, C. UAV reaction detection based on multi-scale feature fusion. In Proceedings of the 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Xi'an, China, 28–30 October 2022; pp. 640–643.
19. Wastupranata, L.M.; Munir, R. UAV Detection using Web Application Approach based on SSD Pre-Trained Model. In Proceedings of the 2021 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), Virtual, 3–4 November 2021; pp. 1–6.
20. Tao, Y.; Zongyang, Z.; Jun, Z.; Xinghua, C.; Fuqiang, Z. Low-altitude small-sized object detection using lightweight feature-enhanced convolutional neural network. *J. Syst. Eng. Electron.* **2021**, *32*, 841–853. [CrossRef]
21. Ye, T.; Zhang, J.; Li, Y.; Zhang, X.; Zhao, Z.; Li, Z. CT-Net: An Efficient Network for Low-Altitude Object Detection Based on Convolution and Transformer. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2507412. [CrossRef]
22. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
23. Ma, J.; Yao, Z.; Xu, C.; Chen, S. Multi-UAV real-time tracking algorithm based on improved PP-YOLO and Deep-SORT. *J. Comput. Appl.* **2022**, *42*, 2885.
24. Li, H.; Yang, J.; Mao, Y.; Hu, Q.; Du, Y.; Peng, J.; Liu, C. A UAV detection algorithm combined with lightweight network. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021; Volume 5, pp. 1865–1872.
25. Liu, Y.; Liu, D.; Wang, B.; Chen, B. Mob-YOLO: A Lightweight UAV Object Detection Method. In Proceedings of the 2022 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD), Harbin, China, 30 November–2 December 2022; pp. 1–6.
26. Liu, R.; Xiao, Y.; Li, Z.; Cao, H. Research on the anti-UAV distributed system for airports: YOLOv5-based auto-targeting device. In Proceedings of the 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China, 20–22 May 2022; pp. 864–867.
27. Xue, S.; Wang, Y.; Lü, Q.; Cao, G. Anti-occlusion target detection algorithm for anti-UAV system based on YOLOX-drone. *Chin. J. Eng.* **2023**, *45*, 1539–1549. [CrossRef]
28. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones* **2023**, *7*, 304. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
32. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. *arXiv* **2022**, arXiv:2208.03641.
33. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
34. Liu, S.; Wang, Y.; Yu, Q.; Liu, H.; Peng, Z. CEAM-YOLOv7: Improved YOLOv7 Based on Channel Expansion and Attention Mechanism for Driver Distraction Behavior Detection. *IEEE Access* **2022**, *10*, 129116–129124. [CrossRef]
35. Zhang, L.; Wang, M.; Liu, K.; Xiao, M.; Wen, Z.; Man, J. An Automatic Fault Detection Method of Freight Train Images Based on BD-YOLO. *IEEE Access* **2022**, *10*, 39613–39626. [CrossRef]
36. Fang, Y.; Guo, X.; Chen, K.; Zhou, Z.; Ye, Q. Accurate and automated detection of surface knots on sawn timbers using YOLO-V5 model. *BioResources* **2021**, *16*, 5390. [CrossRef]

37. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
38. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.