sustainability

MDPI

Article

Aircraft Target Detection from Remote Sensing Images under Complex Meteorological Conditions

Dan Zhong [1,*], Tiehu Li [2], Zhang Pan [3] and Jinxiang Guo [4]

[1] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[2] School of Materials Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China; litiehu@nwpu.edu.cn
[3] The Air Traffic Control Bureau of Civil Aviation Administration of China, Beijing 100022, China; zhp_caac@hotmail.com
[4] The Northwest Air Traffic Control Bureau of Civil Aviation Administration of China, Xi'an 710000, China; gjx_caac2012@hotmail.com
* Correspondence: henryzhongdan@mail.nwpu.edu.cn; Tel.: +86-133-1927-2738

Abstract: Taking all-day, all-weather airport security protection as the application demand, and aiming at the lack of complex meteorological conditions processing capability of current remote sensing image aircraft target detection algorithms, this paper takes the YOLOX algorithm as the basis, reduces model parameters by using depth separable convolution, improves feature extraction speed and detection efficiency, and at the same time, introduces different cavity convolution in its backbone network to increase the perceptual field and improve the model's detection accuracy. Compared with the mainstream target detection algorithms, the proposed YOLOX-DD algorithm has the highest detection accuracy under complex meteorological conditions such as nighttime and dust, and can efficiently and reliably detect the aircraft in other complex meteorological conditions including fog, rain, and snow, with good anti-interference performance.

Keywords: remote sensing images; aircraft target detection; complex meteorological conditions; YOLOX algorithm; depth separable convolution

Citation: Zhong, D.; Li, T.; Pan, Z.; Guo, J. Aircraft Target Detection from Remote Sensing Images under Complex Meteorological Conditions. *Sustainability* **2023**, *15*, 11463. https://doi.org/10.3390/su151411463

Academic Editors: Mohammad Valipour, Xiaoyuan Wang, Junyan Han and Gang Wang

Received: 22 May 2023
Revised: 4 July 2023
Accepted: 21 July 2023
Published: 24 July 2023

1. Introduction

With the development of remote sensing technology, the detection of targets of interest in massive remote sensing images has become an important research area for remote sensing image interpretation. As the core of modern air power acquisition, the efficient detection of aircraft can help to deduce the enemy's military intentions and formulate our operational decisions, achieving the effect of pre-emptive strike. In the civilian sector, remote sensing images of airports and surrounding airspace can be used to detect aircraft, allowing airlines to obtain information on the location and number of aircraft, helping airports to monitor and dispatch aircraft, as well as finding lost aircraft involved in air accidents and playing a role in emergency rescue [1,2].

Traditional detection methods are based on the human mind and visual senses, extracting various attribute features of the aircraft and combining them with template matching or traditional machine learning methods to achieve automatic aircraft target detection [3,4]. However, the proposed features can only superficially characterize aircraft characteristics, which have a low real-time performance when dealing with massive remote sensing data and a poor generalization capability for aircraft detection in complex backgrounds. In recent years, due to the rapid development of deep learning, aircraft detection methods based on deep learning models such as convolutional neural networks for remote sensing images have been intensively researched. These algorithms can be divided into two main categories. The first category is the deep learning target detection method based on candidate regions represented by R-CNN [5–7], which requires the extraction of candidate

regions first through a region selection algorithm before achieving target localization and classification. This type of method has a high localization accuracy, but the detection speed is slightly slower. The other type is based on regression-based deep learning methods, represented by Faster R-CNN [8,9], SSD [10], Center-Net [11], and You Only Look Once (YOLO) [12]. These methods abandon the extraction of candidate regions used in the previous algorithms. Instead, they can directly output the localization and classification results given an input image, making them an end-to-end detection approach. Due to the advantages of the YOLO algorithm's speed and simplicity, many subsequent studies have used it as a reference for improvement, including YOLOv3 [13], YOLOF [14], YOLOX [15], YOLOv8 [16], etc. These algorithms have achieved good detection performance in a variety of application scenarios.

However, the aforementioned algorithms are all oriented toward generic scenarios and suffer from performance losses when addressing specific needs. There are the following reasons: Firstly, in complex weather conditions at airports, targets in images may be affected by low contrast, blurring, and other problems, resulting in changes in the appearance, texture and lighting of targets, making it difficult for methods such as YOLO, FastRCNN and CenterNet to accurately locate and identify targets [17]. Secondly, there are a large number of targets in airports, such as aircraft, vehicles, pedestrians, etc. Occlusion may occur between targets, and generic methods may have limited robustness in handling target occlusion and complex backgrounds, leading to missed or false detections [18]. Thirdly, airport scenarios have targets at different scales, such as large aircraft and small vehicles. The generic method is designed to detect targets at different scales although it uses a multi-scale feature pyramid to detect targets at different scales, but inaccurate detection may occur on targets at extreme scales [19]. In particular, for airport security, it is important to be able to achieve aircraft target detection based on remote sensing images under various meteorological conditions. Therefore, this paper proposes the YOLOX-DD algorithm based on the YOLOX algorithm, which reduces the model parameters by using depth-separable convolution to improve the feature extraction speed and detection efficiency, and introduces a cavity convolution with different cavity rates in the backbone network to increase the perceptual field and improve the detection accuracy of the model. Compared with mainstream target detection algorithms, the proposed algorithm has the highest detection accuracy under complex meteorological conditions such as nighttime and dust, while it can still exclude the influence of interfering pixels and detect field aircraft under other extreme meteorological conditions, with good anti-interference performance. Importantly, the YOLOX-DD algorithm can be applied to the risk detection and collision avoidance scenario in airport scenes, providing robust technological support for the ongoing safety and stability of airport operations.

Next, this paper first introduces the YOLO family of algorithms and YOLOX. Secondly, the paper improves the YOLOX algorithm and designs YOLOX-DD. Furthermore, the paper trains, tests, and validates the performance of the YOLOX-DD algorithm on a self-built dataset. Finally, the paper summarizes the verification results of the YOLOX-DD algorithm and discusses the prospects of its application in the field of civil aviation traffic.

2. Algorithm Introduction

2.1. YOLOX Algorithm

The YOLO v1-v7 algorithms are a series of real-time target detection algorithms that are continuously optimized for accuracy and speed for a variety of scenarios and devices. The YOLOX algorithm is a one-stage target detection algorithm developed by the Megvii Research Institute. Unlike other algorithms in the YOLO series, the YOLOX algorithm converts the YOLO detection head to an anchor-free approach. It also incorporates parallel Decoupled head and SimOTA positive/negative sample allocation strategies, which contribute to its excellent performance in the field of object detection. The YOLOX framework consists of three main components: Backbone, Neck, and Head, as shown in Figure 1.

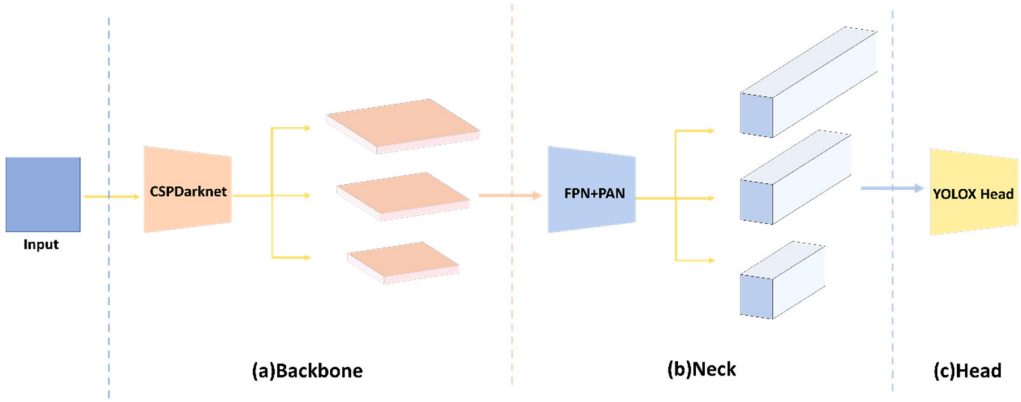

Figure 1. Structure of the YOLOX network.

(1) Input: YOLOX incorporates two types of data augmentation, Mosaic and Mixup, in the network input stage. These techniques enrich the image backgrounds, enhance the model's generalization capability, and improve the robustness of the detection results. The image input model first enters the Focus module, which can downsample the input dimensions without parameters and retains the original image information as much as possible.

(2) Backbone: YOLOX utilizes the CSPDarknet network in the Backbone, primarily through the ConvBNSiLU module and CSPLayer as shown in Figure 1. The ConvBNSiLU module consists of a regular convolution (Conv), batch normalization (BN) processing, and the SiLU activation function. CSPLayer helps reduce the number of parameters and eliminate computation bottlenecks. An SPP structure is added before the last CSPLayer. The SPP structure combines spatial features of different sizes by using Maxpooling layers of different sizes. In YOLOX, the SPP structure employs Maxpooling layers of size (5, 9, 13) to extract features while maintaining the same feature size, thereby enhancing the model's robustness to spatial layout and object variability.

(3) Feature fusion network and prediction module: The Neck part consists of feature pyramid FPN and PAN structure. In YOLOX, the Backbone produces feature maps of sizes $80 \times 80 \times 256$, $40 \times 40 \times 512$, and $20 \times 20 \times 1024$. These feature maps serve as inputs to the FPN structure. FPN facilitates the propagation of deep semantic features to shallower layers, thereby enhancing semantic expression at multiple scales. By transferring information from deeper layers to shallower layers, FPN enriches the semantic representation of the features. The PAN structure performs bottom-up information propagation, allowing strong localization features to be communicated from lower to higher layers. This helps improve the model's ability to localize objects at multiple scales. The PAN structure ensures that both the position and class information of the objects are maximally preserved, enabling accurate object localization.

In Figure 2, the YOLOX Head employs a decoupled head structure. In the Head section, a 1×1 convolution is first used for dimension reduction. In the classification and regression branches, two 3×3 convolutions are utilized. Notably, the classification and regression predictions no longer share parameters. The decoupled head structure takes into account the distinct focuses of classification and localization. By decoupling the parameters for classification and regression tasks, the model can converge faster and achieve higher detection accuracy. This design choice recognizes that the content of interest differs between classification and localization, and it aims to optimize the performance of each task individually.

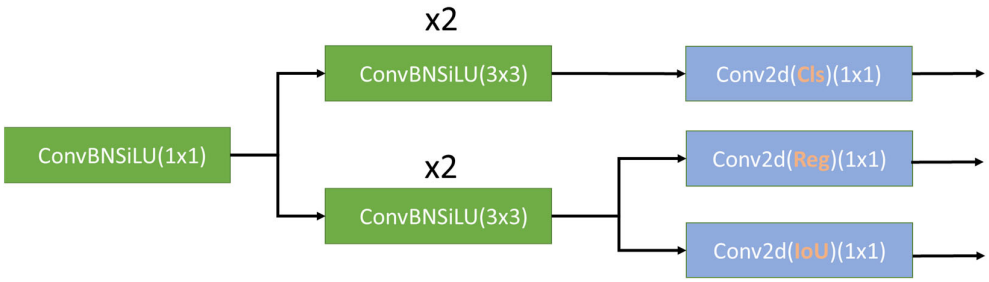

Figure 2. Structure of YOLOX Head.

2.2. YOLOX-DD Algorithm

2.2.1. YOLOX-DD Backbone Network Structure

YOLOX offers multiple network architectures, including YOLOX-S, YOLOX-M, YOLOX-L, and YOLOX-X. Among them, YOLOX-S has the fewest parameters and the fastest runtime. Therefore, in this paper, the ConvBNSiLU module and CSP_Layer in the backbone network of YOLOX-S are improved. The goal is to create a feature extraction network with a wide receptive field, high spatial resolution, and lightweight characteristics by employing depthwise separable convolutions and dilated convolutions. The modified backbone network structure is illustrated in Figure 3.

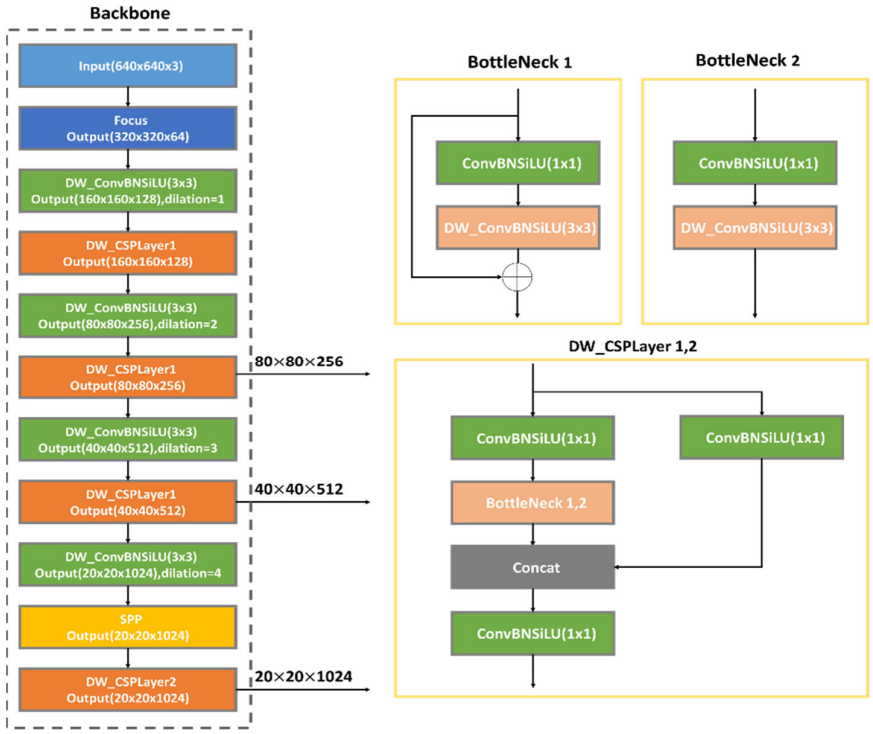

Figure 3. Structure of the YOLOX-DD backbone network.

The classical YOLOX-S network framework, with its many model parameters and large computational effort, consumes more computational resources during training. A large number of convolutional operations are used in the network to extract features, and the presence of a large number of parameters in the convolutional neural network limits the application of the YOLOX-S network to the recognition and detection of field aircraft, and the model parameters and computation should be reduced as much as possible while ensuring the detection accuracy. In this paper, depth-separable convolution is used to achieve the purpose of reducing the number of model parameters, which is an evolution of traditional convolution and can effectively reduce the complexity of operations and

improve the speed of feature extraction and detection efficiency. Also, to improve the detection accuracy of the model, null convolution with different null rates is introduced in the backbone network.

The YOLOX-S backbone network consists of four ConvBNSiLU modules and four CSPLayer modules. In this paper, the focus is on improving these two modules, and the specific approaches are as follows: Firstly, replacing the ordinary convolutions in the four ConvBNSiLU modules with depth-wise separable convolutions. This helps reduce the number of model parameters and computational complexity while maintaining feature extraction capability. Secondly, modifying the BottleNeck in the CSPLayer module. Specifically, replacing the second 3×3 ordinary convolution in the BottleNeck with depth-wise separable convolution. This further reduces the computational cost and enhances the efficiency of the CSPLayer module. Thirdly, to increase the receptive field of the feature maps in the backbone network, different dilation rates are applied to the convolutional operations in the four ConvBNSiLU modules. Specifically, the convolutions in each module use dilation rates of (1,2,3,4) respectively. This promotes better multi-scale feature fusion for subsequent stages like the FPN structure.

The improved feature extraction network has a smaller parameter count, allowing for faster feature extraction. It also has a larger receptive field, which enables more accurate detection of objects at different scales and better adaptation to significant variations in target sizes.

2.2.2. Dilated Convolution

The ConvBNSiLU structure is an important part of the YOLOX-S backbone network. The module uses a 3×3 ordinary convolutional kernel for downsampling, with a wide perceptual field area, but at the expense of spatial resolution, which is too small and can lead to a serious loss of semantic information in the feature map. The semantic information of large-scale objects with large areas and rich features can still be reflected in deeper feature maps, but the semantic information of small-scale objects may disappear completely with the deepening of the network and the target detail information. To ensure that feature details are still enriched without losing resolution, dilated convolution with different hole rates can be used. The principle of null convolution is shown in Figure 4.

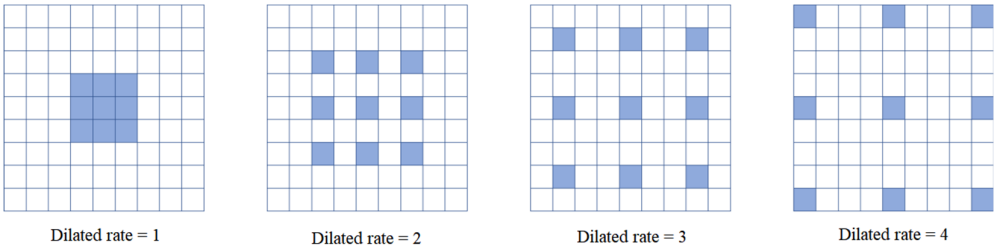

Figure 4. Principle of the Dilated convolution.

Adding the dilated rate to the normal convolution can make the original convolution equivalent to a larger convolution kernel size and expand the perceptual field. Assuming that the size of the convolution kernel of the dilated convolution is k, and the dilated rate is d, the equivalent convolution kernel size is k', which is calculated as follows:

$$k' = k + (k-1)(d-1) \tag{1}$$

Different cavity rates can yield different sizes of sensory fields, and the multi-scale cavity sensory field size is calculated as follows:

$$l_m = l_{m-1} + (k'-1) \times \prod_{i=1}^{m-1} s_i \tag{2}$$

In Equation (2), l_m denotes the receptive field of m layer, l_{m-1} denotes the receptive field of the previous layer, and k' is the equivalent convolutional kernel size of the current layer, and s_i denotes the step size of the i layer.

The algorithm in this paper uses (1,2,3,4) void rates for the ordinary convolution of the four ConvBNSiLU modules in the backbone of the YOLOX-S network, and the sizes of the convolution kernels become 3×3, 5×5, 7×7, and 9×9, respectively, to capture feature information by different perceptual fields.

2.2.3. Depth-Separable Convolution

In order to reduce the number of parameters and save computational resources, depth-separable convolution is used on top of the null convolution. In this paper, we will use depth-separable convolution for the ConvBNSiLU module and CSPLayer layer in the backbone network, and the basic principle is shown in Figure 5.

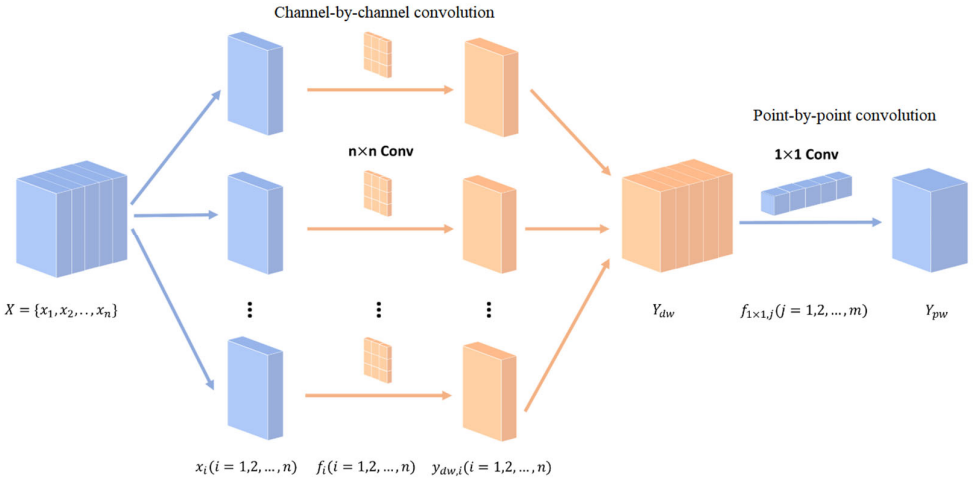

Figure 5. Schematic diagram of depth-separable convolution.

As shown in Figure 5, the depth-separable convolution can be divided into channel-by-channel convolution and point-by-point convolution.

The feature map of the input image is $X = \{x_1, x_2, \ldots, x_n\}$ and n is the number of channels of the input, x_i is the input value of each channel. The convolution kernel for the channel-by-channel convolution is $F_{dw} = \{f_1, f_2, \ldots, f_n\}$. The number of convolution kernels is the same as the number of the input channel n. The output of the channel-by-channel convolution is $Y_{dw} = \{y_{dw,1}, y_{dw,2}, \ldots, y_{dw,n}\}$. $y_{dw,i}$ is calculated as shown in Equation (3),

$$y_{dw,i} = Conv(x_i, f_i)\ (i = 1, 2, \ldots, n) \tag{3}$$

where i is the number of output channels for the channel-by-channel convolution. As can be seen from Equation (3), the output of each channel of the channel-by-channel convolution is related to only one channel of the input, and each channel is operated independently, so that the feature information of different channels at the same spatial location cannot be used effectively, so it is necessary to perform point-by-point convolution to generate a new feature map. The point-by-point convolutional kernel is $F_{pw} = \{f_{1\times1,1}, f_{1\times1,2}, \ldots, f_{1\times1,m}\}$. Point-by-point convolution using a 1×1 convolution kernel, m is the number of convolutional kernels. The point-by-point convolution output is $Y_{pw} = \{y_{pw,1}, y_{pw,2}, \ldots, y_{pw,m}\}$, w is the number of output channels of point-by-point convolution. The same number of convolution kernels as 1×1, $y_{pw,j}$ is calculated as shown in Equation (4):

$$y_{dw,j} = Conv(Y_{dw}, f_{1\times1,j})\ (j = 1, 2, \ldots, m) \tag{4}$$

From Equation (4), the output of each channel of point-by-point convolution is correlated with Y_{dw}. The weighted sum of the values of different channels of Y_{dw} can be calculated by the convolution operation, which well compensates for the deficiency of the channel-by-channel convolution that does not make use of the correlation information between channels.

We suppose the number of input channels of the convolution layer is N, the number of output channels is M, the size of the convolution kernel is $k \times k$, and the number of parameters of the ordinary convolution is:

$$P_{Conv} = k \times k \times N \times M \tag{5}$$

The number of covariates for the deep separable convolution is:

$$P_{DW\_Conv} = k \times k \times N + N \times M \tag{6}$$

The ratio of the number of deeply separable convolutional parameters to the number of ordinary convolutional parameters is:

$$R = \frac{P_{DW\_Conv}}{P_{Conv}} = \frac{1}{M} + \frac{1}{k^2} \tag{7}$$

According to Equation (7), When the number of output channels M is large, the parameter ratio R is about $\frac{1}{k^2}$, and generally $k > 1$. It follows that the depth-separable convolution largely compresses the number of parameters of the convolution process. It can be concluded that depthwise separable convolution greatly compresses the number of parameters in the convolution process. By applying depthwise separable convolution to the ConvBNSiLU modules and CSPLayer layers, the model can be made more lightweight, resulting in efficient feature extraction and improved detection efficiency.

3. Experimental Results and Analysis

The experiments were conducted using the Pytorch deep learning framework, with training and testing carried out on two NVIDIA 2080 Ti graphics cards.

3.1. Surface Aircraft Dataset

In this paper, we use an independently constructed field aircraft dataset with two data categories: airplane and other.

The dataset contains 2007 images of distant aircraft, close aircraft, mutually occluded aircraft, aircraft in different weather scenarios (fog, rain, snow, night, dusty weather) and other targets (airport guidance vehicles, birds, moving vehicles). Among them, 1587 are the training set and 420 are the test set. According to the different scenes in the dataset, aircraft datasets of six special scenes were extracted from the dataset, totaling 523 images (417 images for the training set and 106 images for the test set), as shown in Table 1.

3.2. Network Training

In this paper, we train for the field plane dataset with category 2 and use a stochastic gradient descent optimizer for training. The training batch size is 8, momentum is 0.9, weight decay is 5×10^{-4}, and the maximum number of iterations is 300 epochs. The training learning rate is set to 0.01 at the beginning, and the learning rate is warmed up in the first five epochs, using the exp warm-up curve type to make the model gradually The learning rate is set to 0.01 at the beginning and the learning rate is preheated in the first five epochs. The loss curve of the algorithm YOLOX_DD training in this paper is shown in Figure 6.

Table 1. Special scenario aircraft dataset.

Category	Shade	Fog	Night	Rain	Sand	Snow
Train	126	109	73	39	27	43
Test	28	31	19	10	8	10
Total	154	140	92	49	35	53
Original						
Labeled						

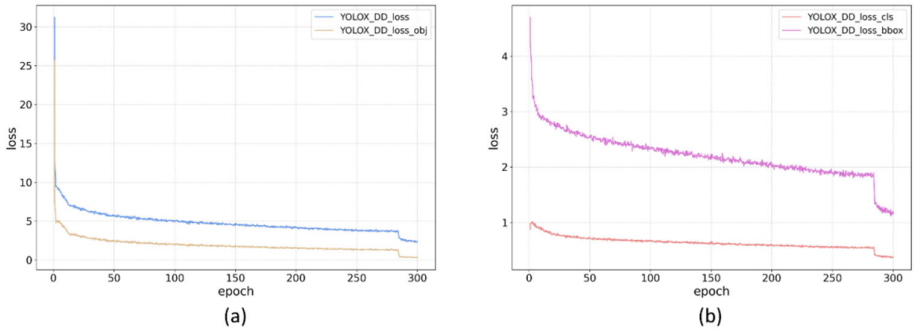

Figure 6. Loss curve of YOLOX-DD.

The losses of the algorithm in this paper are divided into three parts: classification loss (loss_cls), confidence loss (loss_obj), and localization loss (loss_bbox). In Figure 5, the horizontal coordinates represent the number of training iterations, and the vertical coordinates represent the loss values during training. The convergence curves of total loss (loss) and confidence loss are included in Figure 6a, and the convergence curves of localization loss and classification loss are included in Figure 6b.

As can be seen from Figure 6, at the beginning of training, the loss, loss_obj, and loss_bbox of the algorithm in this paper declined relatively rapidly and loss_cls had an upward trend; when the model was iteratively trained up to the 15th epoch, the rate of loss decline gradually became smaller and the four types of loss curves tended to decline smoothly; at the time of training up to 285 epochs. The loss curves all show a rapid decline up to the 285th epoch, and then drop smoothly at the 290th epoch. By the 300th epoch, the loss curves converge, and training is completed.

3.3. Results and Analysis

In order to verify the effectiveness of the improvements made to YOLOX in this paper, the results are shown in Table 2, using YOLOX-DD and YOLOX-S on the test set of the field aircraft dataset and the test set of the special scene aircraft dataset respectively.

YOLOX-DD expands the perceptual field by adding null convolution to the original architecture, enabling the feature maps of different sizes output by the backbone network to capture richer semantic information and enhance the effect of feature fusion. According to Table 2, it can be seen that YOLOX-DD improves the test accuracy by 2.1% on the test set of the field aircraft dataset compared to the original model. The test accuracy on the test set of the special scene aircraft dataset was better than the original model overall, with the improved model improving the test accuracy by 4.6%, 3.4%, 9.5%, and 1.7% in the occlusion, night, sand, and snow scenes respectively. At the same time, YOLOX-DD uses

depth-separable convolution and reduces the parameters of the model by 2.4 M, reducing the number of model parameters and enabling more efficient target detection.

Table 2. Special scenario aircraft inspection test results.

Category	Improved Model AP/%	Original Model AP/%
Surface plane dataset (Test 420)	94.8	92.7
Obscuration (Test 28)	73.1	68.5
Fog (Test 31)	82.6	87
Night (Test 19)	95.5	92.1
Rainy (Test 10)	95.5	88.7
Sandy (Test 8)	97.2	87.7
Snowy (Test 10)	92.3	90.6
Params/M	92.3	8.94

The detection results of the improved algorithm in this paper are shown in Figure 7, mainly for different scenes and one and more aircraft targets. In the absence of occlusion, the detection results of YOLOX-S all show overlapping detection frames, with some of the detection frames being completely contained within another detection frame. In the case of real-time aircraft detection at airports, we do not need overlapping, redundant target detection frames and should obtain as accurate and complete a detection result as possible. In this paper, by adding cavity convolution to YOLOX-S, the perceptual field of the feature map becomes larger and detection results can be obtained with richer feature details. As shown in Figure 7, no redundant detection frames appear in the detection results of YOLOX-DD, and even when the target outline is blurred at night, the edge information of the target can still be extracted, and the target recognition effect is more accurate.

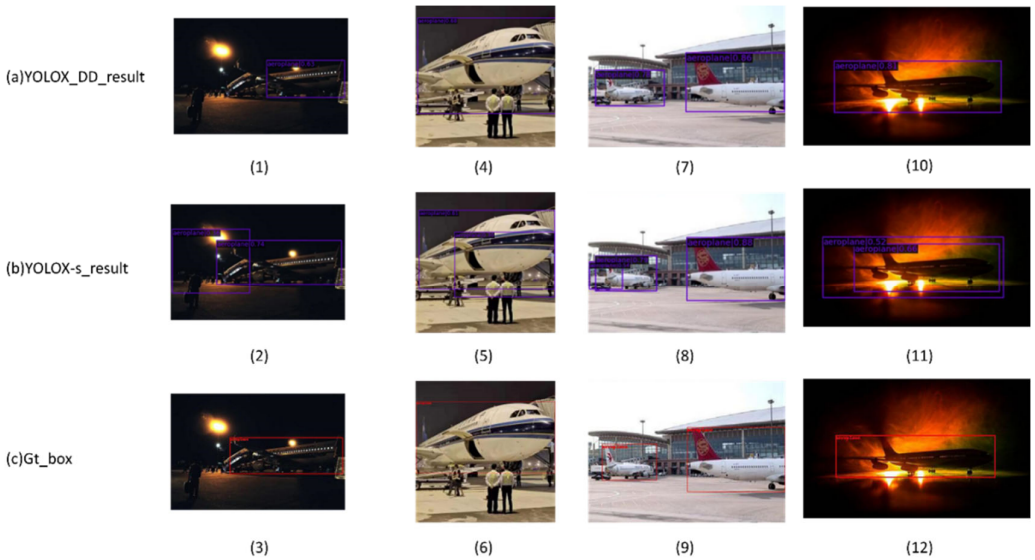

Figure 7. Comparison of aircraft test results.

3.4. Ablation Experiments

In order to demonstrate that each part of the model improvement in this paper has improved the performance of the model, ablation experiments were conducted using the YOLOX-S model as the baseline model, and the aircraft detection accuracy AP (aeroplane), and the number of model parameters (Params) were used in the experiments for comparison. Ablation experiments are an experimental approach to evaluate the impact of these

modifications on model performance by selectively modifying components or parameters in the model. In this paper, three improvements are made to YOLOX: (1) Dila_conv: set the hole rate of (1,2,3,4) for each of the four ConvBNSiLU modules in the backbone network in turn; (2) DW_conv: use depth-separable convolution for each of the four ConvBN-SiLU modules in the backbone network; (3) DW_csp: use depth-separable convolution for each of the CSPLayer in the backbone network using depth-separable convolution, i.e., DW_CSPLayer in Figure 3. Null convolution can increase the perceptual field of feature maps, setting different void rates can promote the effect of multi-scale feature fusion and improve model accuracy, and depth-separable convolution can effectively reduce the complexity of operations and improve the efficiency of model detection.

From Table 3, the YOLOX-S baseline model represents the unimproved YOLOX model, the baseline model aircraft detection accuracy is 92.7%, and the number of model parameters is 8.94 M. From the ablation experiment result 2, it can be seen that the accuracy of using only the Dila_conv model increased by 2%, indicating that using the null convolution can effectively improve the accuracy of the model; from the results 3 and 4, it can be seen that using DW_conv, alone, and DW_csp, respectively, the parameters of the model are reduced by 1.39 M and 1.02 M; from result 5, it can be seen that using DW_conv and DW_csp at the same time, the model parameters are reduced by 2.4 M, indicating that using depth-separable convolution can well reduce the number of parameters of the model; from result 6, it can be seen that using three improvements at the same time, the detection accuracy is improved by 2.1% relative to the baseline model, and the amount of model parameters is reduced. Therefore, the three improvements made in this paper can not only improve the detection performance of the model, but also increase the detection speed of the model.

Table 3. Results of the ablation experiments.

Dila_conv	DW_conv	DW_csp	AP (Aeroplane)/%	Params/M
-	-	-	92.7	8.94
√	-	-	94.7	8.94
-	√	-	92.3	7.55
-	-	√	92.7	7.92
-	√	√	92.9	6.54
√	√	√	94.8	6.54

3.5. Comparison of Different Models

In order to demonstrate the detection effectiveness and superior performance of the improved YOLOX model in this paper, the improved YOLOX-DD model is compared with YOLOv3 [13], Faster-R CNN [8], YOLOF [13], SSD [11], and Center-Net models in this paper, and the above models are tested on the test set of the field aircraft dataset and the test set of the special scene aircraft dataset respectively, and the test results are shown in Tables 4 and 5.

Table 4. Test results for different model field aircraft datasets.

Algorithm Model	AP (Aeroplane)/%	Params/M
YOLOX-DD	94.8	6.54
YOLOv3	84.8 (−10.0)	3.67
Faster-R CNN	86.0 (−8.8)	41.13
Center-Net	83.9 (−10.9)	14.43
YOLOF	91.5 (−3.3)	42.09
SSD	92.6 (−2.2)	23.88

Table 5. Test results for different model-specific scenario aircraft datasets.

Algorithm Model	Shade AP/%	Fog AP/%	Night AP/%	Rain AP/%	Sand AP/%	Snow AP/%
YOLOX_DD	73.1	82.6	95.5	74.4	97.2	92.3
YOLOv3	57.3	78.5	78.3	84.8	76.9	95.1
Faster-R CNN	51.2	67.9	90.0	36.3	76.9	92.3
Center-Net	60.3	77.8	86.8	68.4	80.8	78.5
YOLOF	65.5	80.3	91.1	79.4	78.1	95.7
SSD	67.2	89.5	80.9	79.1	84.3	100.0

As can be seen from Table 4, the model proposed in this paper has more obvious advantages compared with other mainstream target detection algorithms. Compared with YOLOv3, the number of parameters of this paper's model is 2.87 M more than YOLOv3, but the detection accuracy is 10% higher than YOLOv3; the detection accuracy of this paper's model on the field aircraft test set is 8.8% and 10.9% higher than Faster-R CNN and Center-Net respectively, which not only has a better advantage in terms of detection accuracy, but also in terms of the number of parameters. This paper's model is also far less than the other two models; the detection accuracy of YOLOF algorithm and SSD algorithm are 3.3% and 2.2% less than this paper's algorithm respectively, but YOLOF and SSD are far higher than this paper's algorithm in terms of the number of model parameters, the number of parameters of YOLOF is about 6.43 times of this paper's algorithm, and the number of parameters of SSD algorithm is about 3.65 times of this paper's algorithm, thus achieving depth.

As can be seen from Figure 6, this paper's algorithm YOLOX_DD has the highest aircraft detection accuracy in the occlusion, nighttime, and sandy scenes, which is 5.9%, 4.4%, and 12.9% higher than the model ranked 2nd in terms of detection accuracy in the same scenes, respectively. Overall, the YOLOX_DD algorithm has the best detection results in special scenarios. Even under extreme weather conditions, the influence of interfering pixels can still be excluded, and the contour information of the field aircraft can still be detected, which indicates that the algorithm in this paper has better anti-interference performance in terms of detection performance.

4. Conclusions

In this paper, the YOLOX target detection algorithm is applied to the field of civil aviation airport field target detection, and based on the overall framework of YOLOX algorithm, an improved YOLOX-DD algorithm is proposed, which has two main improvements: on the one hand, the model parameters are reduced by using depth-separable convolution, thus improving the feature extraction speed and detection efficiency. On the other hand, the convolution of voids with different void rates is introduced in the backbone network to increase the perceptual field, thus improving the detection accuracy of the model. The improved YOLOX-DD algorithm is trained on a self-constructed field aircraft dataset and experimentally compared with mainstream target detection algorithms. The test finds that the highest detection accuracy is achieved under complex meteorological conditions such as night and dust, while the influence of interfering pixels can still be excluded under other extreme meteorological conditions, and the field aircraft can be detected better with strong robustness. The improved algorithm of YOLOX can be applied to airport field aircraft identification and tracking, remote control tower, airport field video enhancement, and other scenarios. In turn, it is of great significance for civil aviation to ensure the safety of civil aviation operation, improve the operation efficiency, and reduce the occurrence of accidents.

Author Contributions: Conceptualization, D.Z.; Methodology, D.Z.; Project administration, D.Z. and T.L.; Resources, D.Z. and T.L.; Software, D.Z. and Z.P.; Validation, D.Z. and J.G.; Writing-original, D.Z.; Supervision, T.L.; Writing-review & editing, Z.P. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
2. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
3. Wu, Q.; Sun, H.; Sun, X.; Zhang, D.; Fu, K.; Wang, H. Aircraft Recognition in High-Resolution Optical Satellite Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 112–116.
4. Bai, X.; Zhang, H.; Zhou, J. VHR Object Detection Based on Structural Feature Extraction and Query Expansion. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6508–6520.
5. 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] [PubMed]
6. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-Based Fully Convolutional Networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
8. 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.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
9. Zhao, A.; Fu, K.; Sun, H.; Sun, X.; Li, F.; Zhang, D.; Wang, H. An effective method based on acf for aircraft detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 744–748. [CrossRef]
10. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
11. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. Farhadi, A.; Redmon, J. Yolov3: An Incremental Improvement. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1804–2767.
14. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-Level Feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048.
15. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Megvii Technology, YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430v2.
16. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 20 July 2023).
17. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 1792–1800.
18. Li, C.; Qu, Z.; Wang, S.; Liu, L. A method of cross-layer fusion multi-object detection and recognition based on improved faster R-CNN model in complex traffic environment. *Pattern Recognit. Lett.* **2021**, *145*, 127–134. [CrossRef]
19. Zhou, J.; Chen, Z.; Huang, X. Weakly perceived object detection based on an improved CenterNet. *Math. Biosci. Eng.* **2022**, *19*, 12833–12851. [CrossRef] [PubMed]