applied sciences

MDPI

Article

Small-Target Detection Based on an Attention Mechanism for Apron-Monitoring Systems

Hao Liu [1], Meng Ding [1,*], Shuai Li [1], Yubin Xu [2], Shuli Gong [1] and Abdul Nasser Kasule [1]

[1] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; liuhao_lbj@nuaa.edu.cn (H.L.)
[2] China Academy of Civil Aviation Science and Technology, Beijing 100028, China
* Correspondence: nuaa_dm@nuaa.edu.cn

Abstract: Small-target detection suffers from the problems of low average precision and difficulties detecting targets from airport-surface surveillance videos. To address this challenge, this study proposes a small-target detection model based on an attention mechanism. First, a standard airport small-target dataset was established, where the absolute scale of each marked target meets the definition of a small target. Second, using the Mask Scoring R-CNN model as a baseline, an attention module was added to the feature extraction network to enhance its feature representation and improve the accuracy of its small-target detection. A multiscale feature pyramid fusion module was used to fuse more detailed shallow information according to the feature differences of diverse small targets. Finally, a more effective detection branch structure is proposed to improve detection accuracy. Experimental results verify the effectiveness of the proposed method in detecting small targets. Compared to the Mask R-CNN and Mask Scoring R-CNN models, the detection accuracy of the proposed method in two-pixel intervals with the lowest rate of small targets increased by 10%, 3.04% and 16%, 15.15%, respectively. The proposed method proved to have a higher accuracy and be more effective at small-target detection.

Keywords: small targets; Mask Scoring R-CNN; attention mechanism; feature fusion

check for updates

Citation: Liu, H.; Ding, M.; Li, S.; Xu, Y.; Gong, S.; Kasule, A.N. Small-Target Detection Based on an Attention Mechanism for Apron-Monitoring Systems. *Appl. Sci.* **2023**, *13*, 5231. https://doi.org/10.3390/app13095231

Academic Editor: Andrea Prati

Received: 22 March 2023
Revised: 19 April 2023
Accepted: 20 April 2023
Published: 22 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

If an aircraft encounters safety problems during a flight, the consequences are severe [1]. Ensuring airport safety, in addition to flight safety, is particularly important. When taxiing and parking an aircraft, it is necessary to frequently check whether there are other personnel on the airport surface. This is complicated by the fact that the appearance of staff and pedestrians on the airport surface is random and irregular. Therefore, real-time monitoring of staff and pedestrians is challenging [2]. Furthermore, staff and pedestrians can be regarded as small targets relative to the aircraft [3]. Owing to problems such as light reflection and long distances, small targets are difficult to detect [4]. It is necessary to ensure that there are no pedestrians or other obstacles on the apron and flight region during the taxiing and parking of an aircraft [5].

In response to the first problem, the randomness of airport personnel being present on the airport surface, the detection of targets on the apron and flight region relies mainly on monitoring and surveillance systems [4]. A vision-based surveillance system can effectively reduce labour costs and work pressure on supervisors. Moreover, a vision-based surveillance system produces 7 × 24 h of uninterrupted intelligent patrol, which improves patrol efficiency and alarm accuracy, effectively solving the problem of the random appearance of airport personnel. In addition, vision-based surveillance systems can be used to modify traditional security operations and management modes as well as improve the value of monitoring equipment and the efficiency of decision-making [6]. Finally, they provide off-site and intelligent management methods for airport flight area

operation specifications, quality control, and safety precautions, and improve precise supervision and scientific safety operation management capabilities [7].

However, the second problem, or the difficulties associated with small-target detection, arise from the following: (1) the automatic dependent surveillance-broadcast (ADS-B) [8,9] cannot detect non-cooperating targets, such as pedestrians and staff, and (2) the low resolution of the surface movement radar (SMR) [10,11] leads to the inability to effectively detect small targets. Although vision-based surveillance systems can monitor all targets on the entire apron in real-time and can even detect some small targets, numerous problems are still encountered in the detection of small targets, such as during the fine segmentation and location of such targets and when learning their non-salient features [12]; further, bad weather can result in the omission of personnel or obstacles on the apron. In summary, the above difficulties limit the detection accuracy of small targets in vision-based surveillance systems. This paper proposes a method for small-target detection on the apron to address the aforementioned problem. The contributions of this study can be summarised as follows:

(1) A small-target detection model that can be applied to apron monitoring systems is proposed to monitor small targets (staff and pedestrians) on the airport surface.
(2) To enhance small-target detection accuracy, an improved small-target detection method is proposed to better extract and fuse features.
(3) A new, standard apron small-target dataset, mainly composed of airport apron pedestrians and staff, is established.
(4) The effectiveness and feasibility of the proposed method are verified on a real apron, and small targets are effectively detected, proving that the proposed method can be applied to monitor a real apron.

The remainder of this paper is organised as follows: Section 2 presents a brief review of recent methods related to small-target detection. The proposed method is described in detail in Section 3. Section 4 presents the experimental results and analysis, including a comparison of the results with those of the current small-target detection algorithm. Finally, in Section 5, we present our conclusions.

2. Related Works

Typical airport surface monitoring systems mainly monitor pedestrians, staff, and vehicles on the airport surface, but it must be noted that people are much smaller in size than cars; therefore, small-target surveillance on the airport surface should focus on monitoring people. Owing to the large area of the apron, it takes a significant amount of time for staff to monitor targets on the apron, and sometimes, owing to bad weather, personnel or obstacles on the airport surface may get omitted. If not detected in time, such omissions may lead to serious safety hazards. Additionally, the limitations of ADS-B and SMR make it particularly important to develop a small-target surveillance system. Based on these criteria, vision-based surveillance systems have addressed the issues.

In recent years, deep convolutional neural networks have been widely used in various vision-based surveillance methods, especially for the detection of small targets. Furthermore, the detection of small targets is widely used in various fields, such as infrared small-target detection [13], defect detection for industry 4.0 [14], and pest detection in smart farming [15]. Such detection methods are divided into two categories according to the processing of candidate frames [16]: (1) One-stage target detection methods based on regression, which take the entire image as an input to increase the receptive field of the targets on the image and regress the position and category information of the targets at different positions on the image. The most representative methods of this category are the YOLO [17,18] series, SSD [19], and so on. (2) In the two-stage target detection method, the target candidate frame that may exist on the image is first extracted, the candidate frame of each region is classified, and then position regression is performed. Such methods mainly include the R-CNN [20], Fast R-CNN [21], Faster R-CNN [22], and Mask R-CNN [23–25]. The first group of methods (one-stage methods) have fast detection speeds and a high adaptability to large targets; however, small targets are easily missed. The second category

of methods (two-stage methods) exhibit a relatively high accuracy in small-target detection and are therefore highly suitable for this task. Thus, we chose the Mask R-CNN as our baseline.

However, numerous challenges are encountered in the detection of small targets when using deep learning. First, target detection based on deep learning uses a convolutional neural network (CNN) as a feature extraction tool. When attempting to increase the receptive field in the CNN, the feature map shrinks, and the stride length may be larger than the size of the small targets; this makes it difficult for the small target to be passed on to the deep network during network convolution [26]. Second, in commonly used datasets, such as Microsoft Common Objects in Context (MS COCO), the sample size of small targets accounts for a small proportion of the total targets, and the size difference between large and small targets is relatively significant. This leads to certain difficulties during target-aware network adaptation. Third, owing to the complexity of the airport surface, it is highly difficult for staff to monitor the targets of the apron. Target detection is the basic module of vision-based surveillance and is directly related to the overall performance of the entire surveillance system. However, because the apron targets are too small, they have few pixels, and their shapes and their outlines are unclear. The small targets appear similar to the surrounding background.

This study improves on the Mask Scoring R-CNN algorithm and proposes a method of small-target detection based on an attention mechanism, as shown in Figure 1. The proposed method can not only obtain the location and category of the small targets, but also finely segment them to obtain the corresponding geometric properties (including the length, width, area, contour, centre, etc., of the targets).

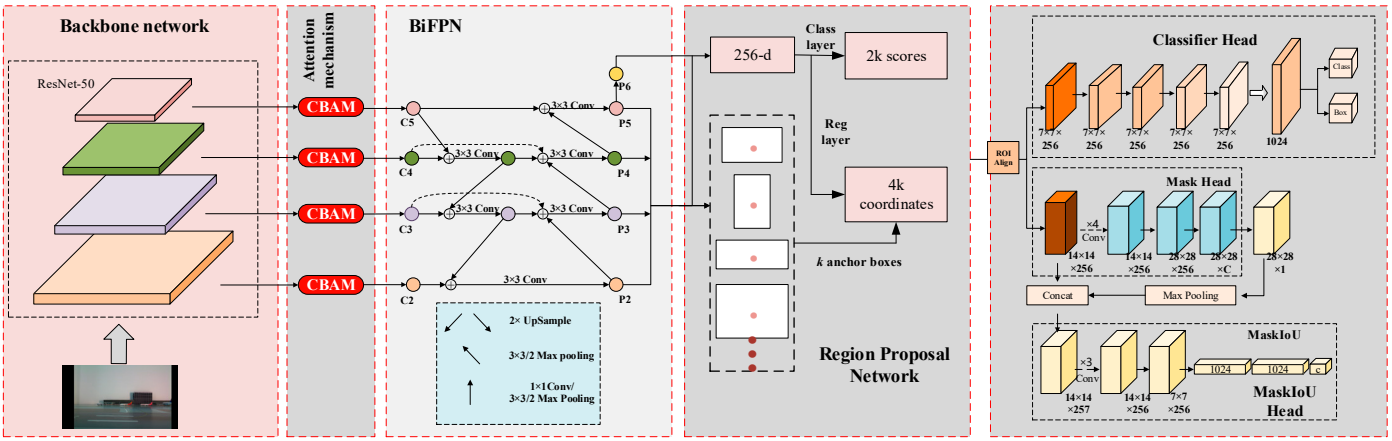


Figure 1. Network structure.

3. Materials and Methods

This study presents a model that shows improvements in terms of three aspects: feature extraction, feature fusion, and classification. First, to enable the network to learn to extract the representative features of small targets on airport aprons and provide more effective feature information for the classifier and mask prediction, the proposed method adds an attention module [27–29] to the feature-extraction network. The dependency relationship and positional relationship in the feature space guide the network to increase the weight of useful features of small targets on airport surfaces, making the network pay more attention to features that are conducive to the detection of small targets and ignore redundant and invalid information. Second, for the feature fusion module, a bidirectional feature pyramid network (BiFPN) is introduced [30,31] and then compared with the traditional feature pyramid network (FPN). BiFPN uses weighted fusion that enables the network to perform fusion more effectively. Finally, a more effective detection branch structure is proposed by changing the network structure in the original algorithm, thereby further improving the detection accuracy of small targets [32,33].

3.1. Attention Module: CBAM

Spatially, small targets on airport aprons are randomly located, and the attention module can guide the neural network to highlight important features during detection to improve the locating accuracy. Therefore, the algorithm used in this study refers to the convolutional block attention module (CBAM) [34,35]. The differences in the attention module (Figure 2) between the original method and the proposed method are as follows: the attention module of the original method is embedded into the feature extraction network, which is a connection module of residual blocks; however, this study uses an independent attention module to connect the attention module with the output of the feature extraction network, the proposed method does not change the structure of the original feature extraction network, which can be initialised with pre-trained weights. Each attention module comprises two sub-modules: channel attention (Figure 3) and spatial attention modules (Figure 4).

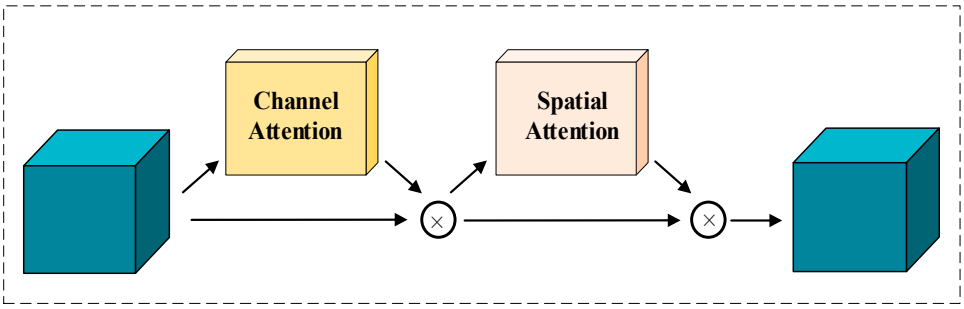


Figure 2. Attention module.

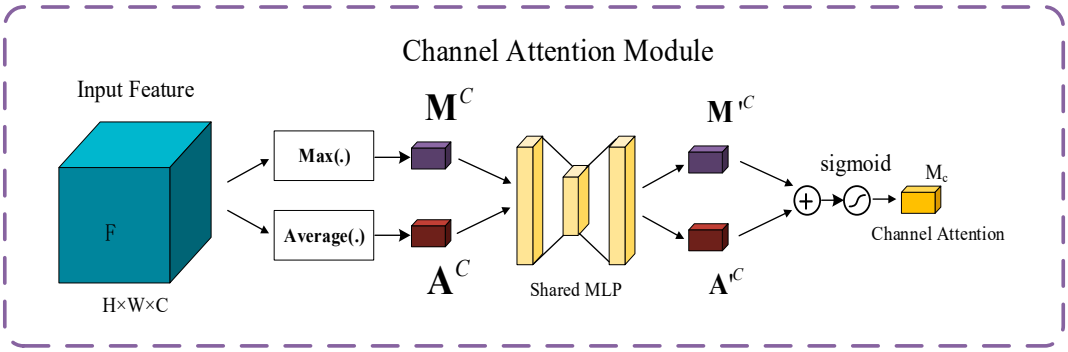


Figure 3. Channel attention module.

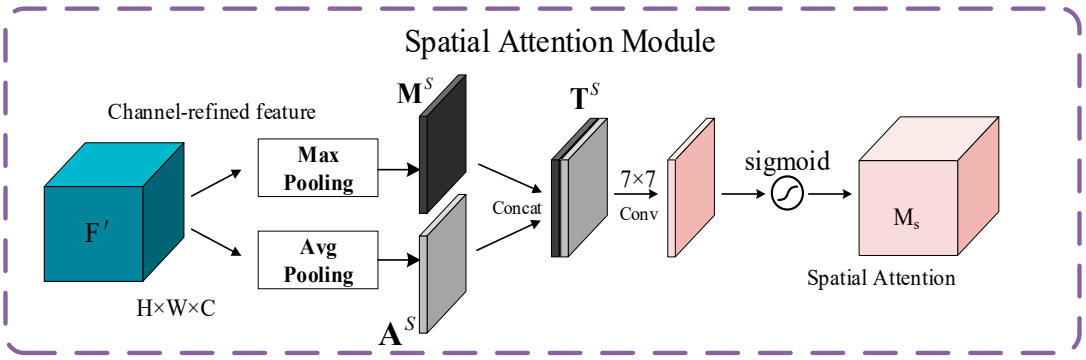


Figure 4. Spatial attention module.

1. Channel attention module: In the convolutional neural network, the size of the feature map slowly decreases with an increase in the depth of the network, and the number of feature channels increases exponentially as a result. The information extracted from

one feature channel differs from that extracted from the other, and more features can be extracted using a conventional neural network. Although it is necessary to increase the number of feature channels in the feature extraction process, not all features are necessary, and some features only play an auxiliary role. The purpose of the channel attention module is to allow the network to give greater weight to the important feature channels, amplify feature information that contributes significantly to subsequent tasks, and suppress the irrelevant feature channels. The feature map processed via channel attention pays more attention to the amplified feature information.

Different feature channels contain different convolutional information. The channel attention module allows the network to place more weight on the important small-target feature channels, thereby suppressing insignificant feature channels, as shown in Figure 3. Global maximum pooling and global average pooling are performed on the input feature map to aggregate the spatial information on the feature map, as can be seen from this figure. Subsequently, two different average pooling vectors and maximum pooling vectors are generated from the previous step. Both vectors are separately sent to the shared multilayer perceptron (MLP). To reduce the parameter overhead, the number of neurones in the first layer is C/r (r is the reduction rate of the attention channel) using rectified linear unit (ReLU) activation, and the second layer uses C neurones. Therefore, the length of the feature vector is first decreased and then increased, to reduce certain parameter overheads and improve the speed of attention generation. After MLP processing, the output feature vectors are added to form a feature vector via element-wise addition [36–38]. The output features are then added together and activated by the sigmoid to generate the final channel attention features. The channel attention can be calculated as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \quad (1)$$

where F represents the input feature map, F^c_{avg} represents average pooled feature vectors, F^c_{max} represents the maximum pooling vectors, and W_0 and W_1 represent the weight parameter of the shared MLP, σ which represents the sigmoid activation function.

2. *Spatial attention module:* In addition to knowing which features of the channel are important, the feature extraction network can distinguish the attention of local spatial features. The spatial attention module guides the network to highlight areas containing useful information based on spatial relationships, and small targets may exist in specific areas. In most cases, owing to the random location of small targets on the apron, highlighting important local feature information is conducive to positioning these targets. The input of the spatial attention module is the feature from the output of the channel attention module. In the channel dimension, two spatial two-dimensional attention maps are obtained using global average pooling and global maximum pooling, also called average pooling and maximum pooling. The two maps are connected using channel concatenation and then a 7×7 convolution operation is performed. Spatial feature attention is generated when the sigmoid function is activated [39–41]. Spatial attention is performed as follows:

$$\begin{aligned} M_s(F) &= \sigma(C^{7\times7}(cat[AvgPool(F), MaxPool(F)])) \\ &= \sigma(C^{7\times7}(cat[F^s_{avg}, F^s_{max}])) \end{aligned} \quad (2)$$

where σ represents the sigmoid activation function, $C^{7\times7}$ represents the convolution with the kernel of 7×7, and *cat* represents the splicing of the extended channel direction.

3.2. Feature Fusion: BiFPN

The smallest feature layer is 32 times smaller than the original image, which is not conducive to the detection of small targets on the apron. When the area of small targets is less than 32^2, the extracted features are compressed to less than one pixel, essentially losing the features of the small targets. Deep convolutional layers contain more semantic information about small targets; this helps with target classification. Shallow convolutional

layers are more likely to be activated by local textures and can capture more specific details, which is suitable for detecting small targets. Therefore, combining both layers improves the detection of small targets.

Different levels of features are directly added when performing feature fusion in an FPN; this operation causes different levels of features to have the same weight. This, in turn, ensures that different feature layers contribute equally to the detection results. By adding a top-down path, shallow features can fuse powerful semantic information with deep features. However, because of the relatively large number of convolutions from the bottom to the top layer, dozens or even hundreds of convolution layers are required; thus, many detailed features of the target are lost in the process. Therefore, a lack of more specific details on deep features affects the detection accuracy of the network. BiFPN adds low-to-top channels and weighted fusion, which compensates for the lack of detailed information and produces more effective multiscale feature fusion.

In Figure 5, C_i and f_i represent the features output by the feature extraction network and the fused feature, respectively. After C_{i+1} is up-sampled, it is fused with C_i according to (3), and then the intermediate feature f_i^{tmp} is obtained through 3×3 convolution. Then the same method is used to fuse C_i, f_i^{tmp} and f_{i-1} to generate f_i. The specific formula is shown in (4).

$$F = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} \cdot I_i \tag{3}$$

$$\begin{cases} f_i^{tmp} = conv\left(\frac{\omega_1 \cdot C_i + \omega_2 \cdot up(C_{i+1})}{\omega_1 + \omega_2 + \varepsilon}\right) \\ f_i = conv\left(\frac{\omega_1' \cdot C_i + \omega_2' \cdot N_i^{tmp} + \omega_3' \cdot down(N_{i-1})}{\omega_1' + \omega_2' + \omega_3' + \varepsilon}\right) \end{cases} \tag{4}$$

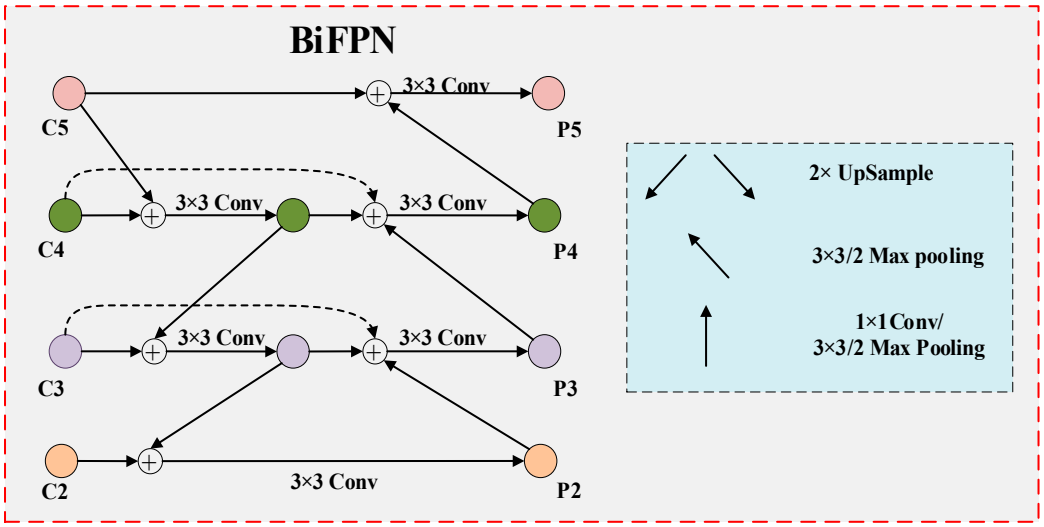


Figure 5. The structure of BiFPN.

3.3. Classifier Head

The mask head of the Mask Scoring R-CNN method requires a classifier head to provide a category and bounding box for small targets, where the bounding box is used to clip the corresponding feature layer. It is then sent to the network for segmentation, and the final target mask is selected according to the target category. During detection, the original target detector predicts numerous target frames containing small targets; however, these target frames are redundant, and some target frames highly overlap. Therefore, numerous duplicated boxes must be filtered out in a nonmaximal, suppressed manner. We believe that the input feature involves numerous bounding box information because it is obtained through dozens of convolution layers. In this study, we adopted a deeper network structure (Figure 6). Four convolutions were used to replace the first fully connected layer to recode

the input features. To weaken the influence of external information during detection and highlight the target features, the output size of each convolution layer is always consistent with the input [42].

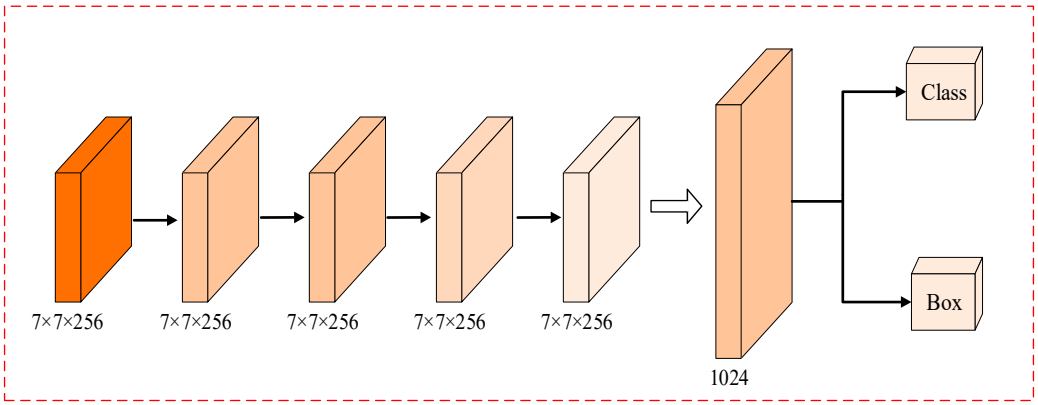


Figure 6. Classifier head.

In this study, the number of categories of airport small targets and other daily detection tasks was significantly lower than the number of common targets in the COCO dataset, which contains 80 types of targets. This study only detects and locates people. Further, using only the fully connected layer to complete two tasks is obviously not conducive to the positioning of small targets on the apron. Therefore, given that the classification task is not complex, and considering that the use of fully connected and convolutional branches greatly increases the network parameters and computational consumption, we use a convolutional layer that is more suitable for localisation to complete the classification and localisation tasks. As mentioned above, the network was designed to follow the setup of R-CNN, with four layers of shared convolution and one fully connected layer. The convolutional layer was used to extract local information suitable for classifying and locating targets in the proposed box, and the full connection introduced position coding, integrating global information through different position information. At the network parameter level, the number of parameters of the original network full connection was $14 \times 14 \times 256 \times 1024 = 51{,}380{,}224$, whereas the parameter quantity of the convolutional layer was only $3 \times 3 \times 256 \times 256 \times 4 = 2{,}359{,}291$, which was significantly smaller than the number of parameters of the full connection [43].

4. Experimental Results

4.1. Dataset

Because no datasets on airport small targets are publicly available, we could not compare our results with those reported in other studies under the same standard. In this regard, we created a small target dataset simulating the airport and applied our method using this dataset. We also performed several experiments using this dataset.

The sizes of the small targets used in this experiment were based on the definition of relative scale. The width and height of the targets are generally considered to be less than one-tenth of the image, and the specific area is generally 32×32 pixels. In other words, if the area covered by the target is less than 0.25% of the entire image, it can be considered a small target. We took 300 high-quality images of an airport surface, which included 1500 small targets, by constructing a simulated airport experimental platform (Figure 7). All the images were calibrated pixel-by-pixel using LabelMe, and the dataset was split in a 7:3 ratio between the training set and test set. The format of our dataset was the same as that of the COCO dataset. According to the pixel size statistics of the targets, the maximum pixel was 45×21, and the maximum proportion of the target size was 0.12%; therefore, all targets were small. Owing to the different pixel proportions of different targets, we divided the small targets into 11 proportional intervals from 0 to 0.25%, and the number of

correctly identified targets in each interval was computed to compare the correct rate of identification using different methods. The target number of statistics for each interval was as shown in Figure 8:

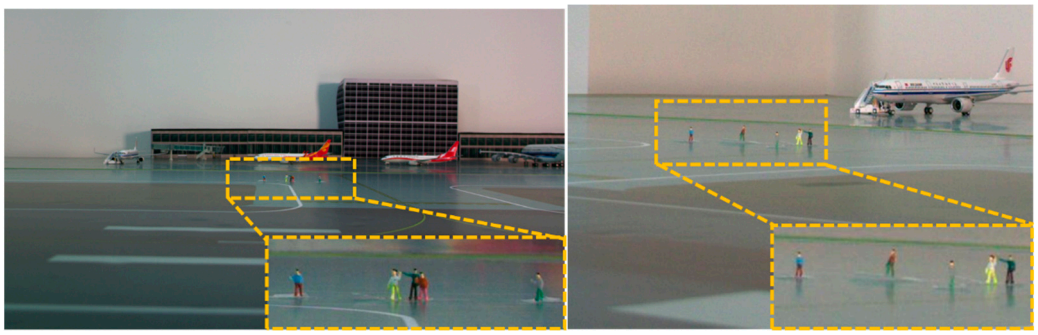


Figure 7. Airport small-target dataset.

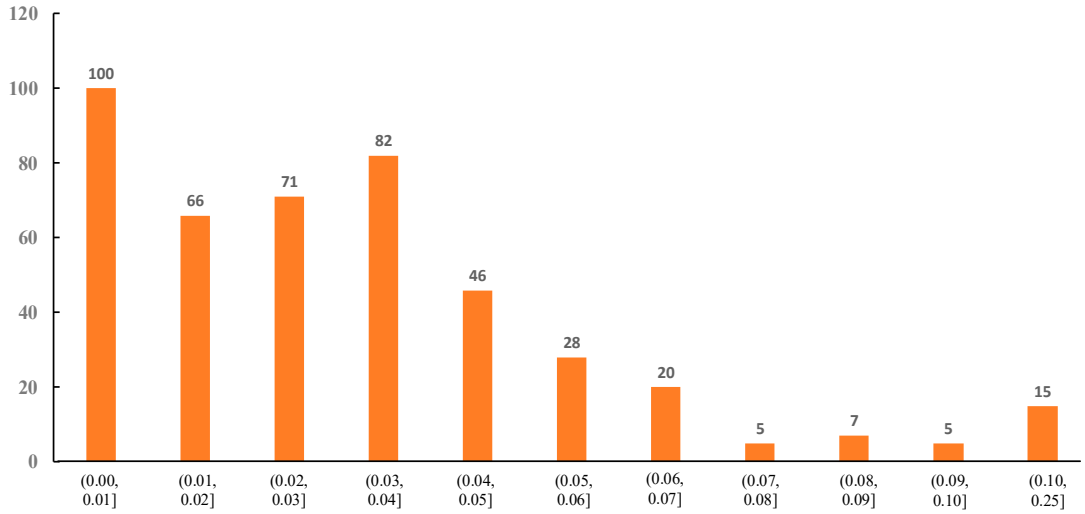


Figure 8. Dataset details.

4.2. Experimental Details

We used the Mask Scoring R-CNN model as the baseline in our comparison experiments. All the pre-trained models used in this study are publicly available, and the accuracy of the different pixel ranges was taken as the evaluation index. All the models were trained on a GPU with two images per batch, and the network weights were updated using the Adam optimiser. Unless otherwise specified, all the models were trained for 96 epochs with an initial learning rate of 0.001, and the policy was updated with a cosine cooling learning rate. The minimum learning rate was set to 0.001 times the initial learning rate, and the models were initialised with ImageNet pre-trained weights. A multiscale training strategy was added to the final model, and the short side of the input image was randomly sampled between 640 and 768 with a step size of 32.

4.3. Ablation Experiments

According to the pixel size of all targets in the dataset, in which the maximum pixel was 45 × 21 and the maximum proportion of the target size was 0.12%, the number of correctly identified targets in 11 proportional intervals was counted. The correct rate of identification using different methods was compared. The results of the ablation experiment are as follows.

Attention module-CBAM: In comparing Tables 1 and 2, it can be seen that the detection results of the attention mechanism are better than the baseline, which were increased by 8% in the range of (0, 0.01) and 6.06% in the range of (0.01, 0.02), and the results of other ranges

reach 100%, which shows that the attention mechanism added in this study had a more noticeable effect and is conducive to the detection of small targets.

Table 1. Target-recognition results of Mask Scoring R-CNN.

Interval (%)	Targets	Correct Number	Correct Rate
(0, 0.01)	100	7	7%
(0.01, 0.02)	66	55	83.33%
(0.02, 0.03)	71	70	98.59%
(0.03, 0.04)	82	81	98.78%
(0.04, 0.05)	46	46	100%
(0.05, 0.06)	28	28	100%
(0.06, 0.07)	20	20	100%
(0.07, 0.08)	5	5	100%
(0.08, 0.09)	7	7	100%
(0.09, 0.10)	5	5	100%
(0.1, ~0.25)	15	15	100%

Table 2. Target-recognition results of CBAM.

Interval (%)	Targets	Correct Number	Correct Rate
(0, 0.01)	100	15	15%
(0.01, 0.02)	66	59	89.39%
(0.02, 0.03)	71	71	100%
(0.03, 0.04)	82	82	100%
(0.04, 0.05)	46	46	100%
(0.05, 0.06)	28	28	100%
(0.06, 0.07)	20	20	100%
(0.07, 0.08)	5	5	100%
(0.08, 0.09)	7	7	100%
(0.09, 0.10)	5	5	100%
(0.1, ~0.25)	15	15	100%

Feature fusion-BiFPN: It can be seen from Table 3 that adding the feature fusion module BiFPN alone, compared with Mask Scoring R-CNN from the experimental results, increased the detection accuracy in the range of (0, 0.01) by 3%, and the detection accuracy in the range of (0.01, 0.02) was increased by 4.55%. Overall, there were seven more small targets detected than in the baseline.

Classifier head: From Table 4, it is visible that the detection head added in this study also had noticeable effects. The detection accuracy in the range of (0, 0.01) was improved by 2%, in (0.01, 0.02) by 7.58%, and the total number of detections was increased by nine small targets.

Table 3. Target-recognition results of BiFPN.

Interval (%)	Targets	Correct Number	Correct Rate
(0, 0.01)	100	10	10%
(0.01, 0.02)	66	58	87.88%
(0.02, 0.03)	71	70	98.59%
(0.03, 0.04)	82	82	100%
(0.04, 0.05)	46	46	100%
(0.05, 0.06)	28	28	100%
(0.06, 0.07)	20	20	100%
(0.07, 0.08)	5	5	100%
(0.08, 0.09)	7	7	100%
(0.09, 0.10)	5	5	100%
(0.1, ~0.25)	15	15	100%

Table 4. Target-recognition results of Classifier head.

Interval (%)	Targets	Correct Number	Correct Rate
(0, 0.01)	100	9	9%
(0.01, 0.02)	66	60	90.91%
(0.02, 0.03)	71	71	100%
(0.03, 0.04)	82	82	100%
(0.04, 0.05)	46	46	100%
(0.05, 0.06)	28	28	100%
(0.06, 0.07)	20	20	100%
(0.07, 0.08)	5	5	100%
(0.08, 0.09)	7	7	100%
(0.09, 0.10)	5	5	100%
(0.1, ~0.25)	15	15	100%

Our method: As shown in Table 5, the proposed method, or adding the above three in specific positions, obtained the best results. The range of (0, 0.01) increased by 16%, the range of (0.01, 0.02) increased by 15.15%, and the overall number of small targets detected increased by 27, which effectively proves that the proposed method better detects small targets.

According to the results, the detection method's accuracy was improved due to the use of the Mask Scoring R-CNN. When only the CBAM was included, the detection performance was the most improved (Figure 9), compared with the baseline in the scale range of (0, 0.01). That is, the detection accuracy was improved by 8%. Compared with the baseline, the method proposed in this study was improved 16% and 15.15% in the range of (0, 0.01) and (0.01, 0.02), respectively; all other sections were also improved. As can be seen in Table 6, the model parameters exhibit certain changes while the remaining changes are minor. This is sufficient to prove that the proposed method has a higher detection accuracy for small targets.

Table 5. Target-recognition results of our method.

Interval (%)	Targets	Correct Number	Correct Rate
(0, 0.01)	100	23	23%
(0.01, 0.02)	66	65	98.48%
(0.02, 0.03)	71	71	100%
(0.03, 0.04)	82	82	100%
(0.04, 0.05)	46	46	100%
(0.05, 0.06)	28	28	100%
(0.06, 0.07)	20	20	100%
(0.07, 0.08)	5	5	100%
(0.08, 0.09)	7	7	100%
(0.09, 0.10)	5	5	100%
(0.1, ~0.25)	15	15	100%

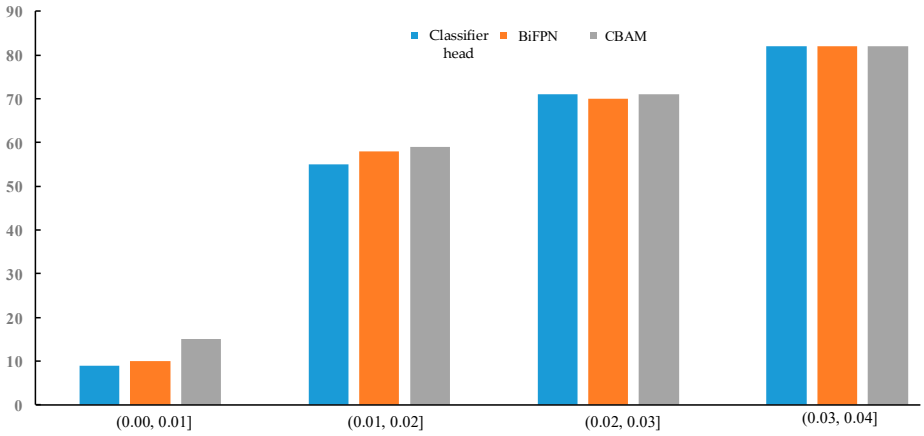


Figure 9. Results of different modules.

Table 6. Results of metrics.

Method	Parameters	Training Time	Memory
Baseline	60.01 M	1.7 h	5264 MB
Baseline + CBAM	61.75 M	1.7 h	5302 MB
Baseline + classifier head	59.32 M	1.6 h	5232 MB
Baseline + BiFPN	96.31 M	2.1 h	6203 MB
The proposed	98.32 M	2.2 h	6283 MB

4.4. Comparison with Other Methods

Compared to the Mask R-CNN experiment, our experimental results contained 17 pictures with undetectable small targets, while the Mask R-CNN experimental results contained 24 pictures with undetectable small targets. The statistics of the Mask R-CNN detection results are shown in Figure 8. As shown in Figure 10, the comparison shows that in the (0, 0.01) interval, the accuracy of our results was 23%, whereas that of the Mask R-CNN was 10%; further, the accuracy of the results obtained by the Mask Scoring R-CNN was 7%. When the proportion of small targets was in the (0.01, 0.02) interval, the accuracy rate of our experiment was 98.48%, while the accuracy of Mask R-CNN detection results was 95.45% and the correct rate of Mask Scoring R-CNN detection results was 83.33%, and the correct rate in other proportions was 100%. Therefore, the detection results obtained by this method are much higher than those obtained using the traditional small-target detection method.

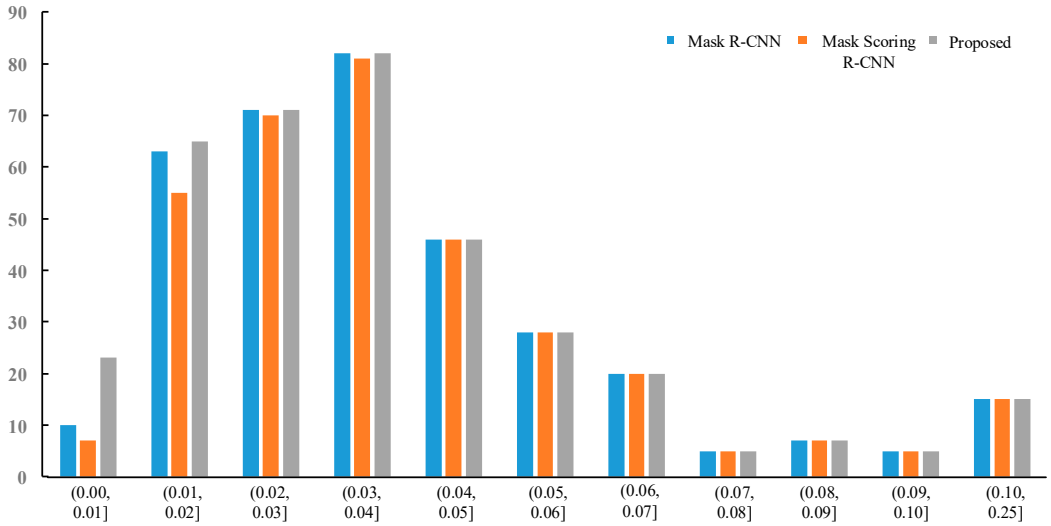
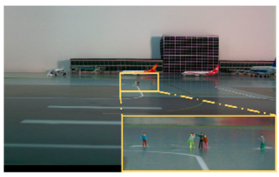
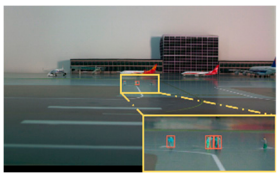
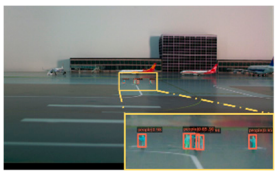
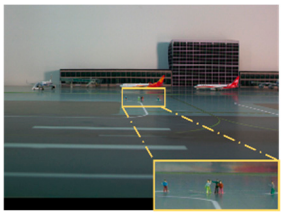
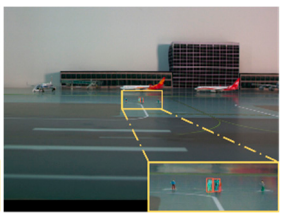
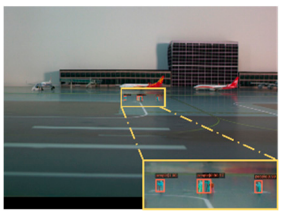
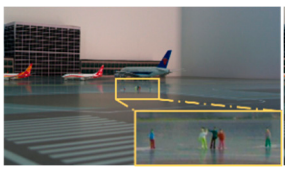
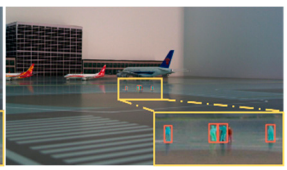
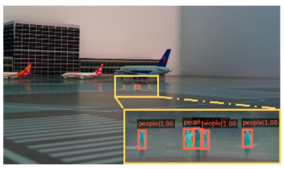
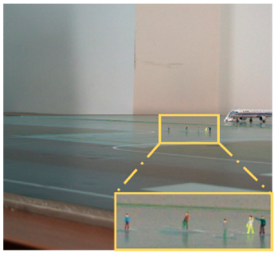
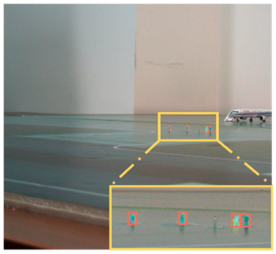
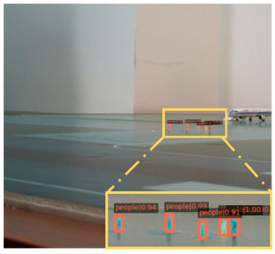


Figure 10. Results of the quantitative comparison.

Qualitative comparison: As shown in Table 7,it can be seen that the four images of the detection results of the Mask R-CNN missed the targets that were detected by the proposed method. From (a)–(d), it is observed that our method detected all the staff on the airport surface. Therefore, the detection effectiveness of our method is better because it accurately and efficiently detected all small targets.

Table 7. Target-recognition results of qualitative comparison.

No.	Input Image	Mask R-CNN	Ours
(a)			
(b)			
(c)			
(d)			

To verify the effectiveness of the proposed method, it was applied to a real airport surface to detect apron personnel. All staff in the figures meet the definition of small targets. As can be seen in Figure 11, all the staff were detected with high accuracy, which proves that the proposed method has a desirable effect on the detection accuracy of small targets on the airport apron. Note that the left side presents the original figure, and the image detected by the proposed method is presented on the right side.

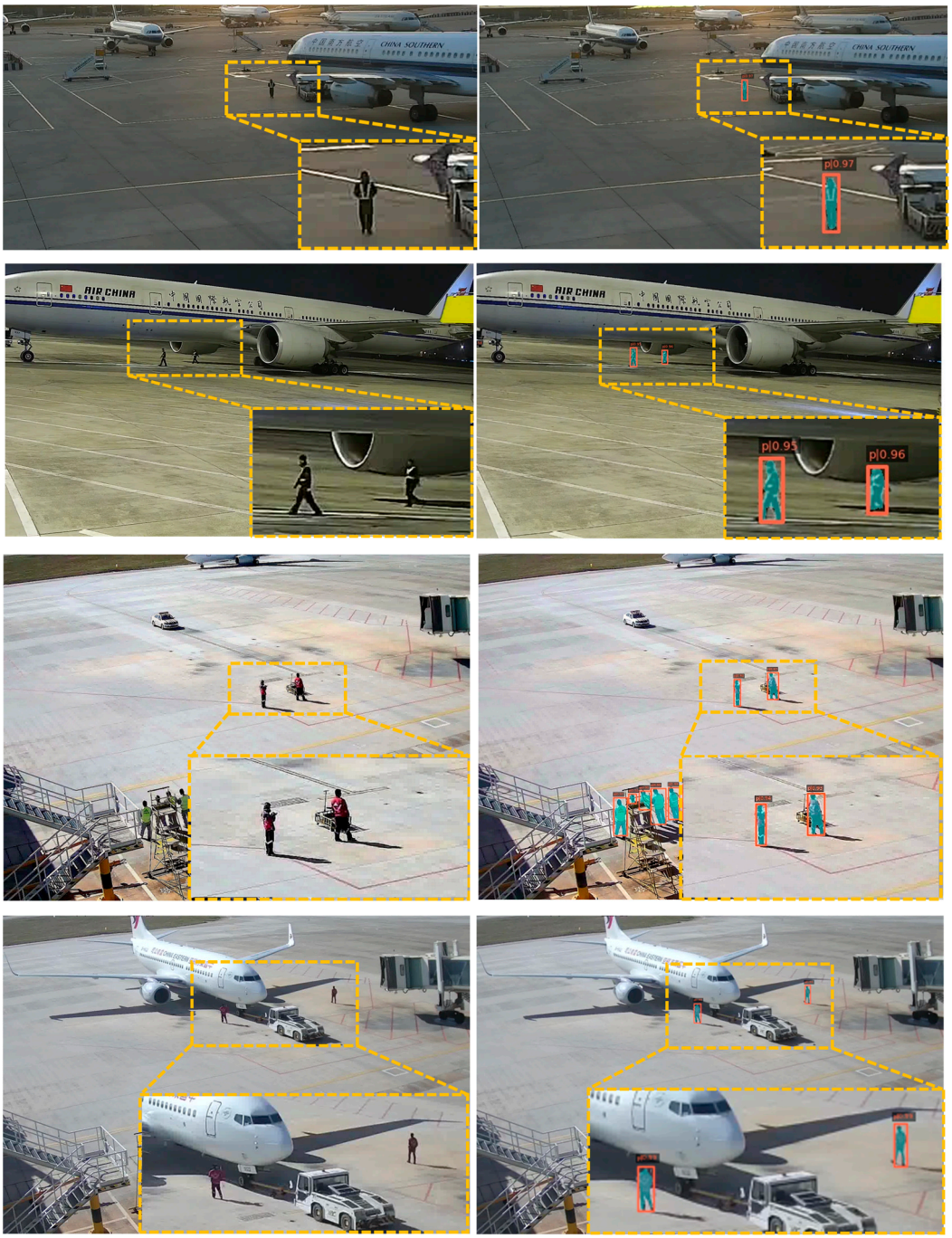


Figure 11. Real apron detection results.

5. Conclusions

This study proposes a small-target detection method based on an attention mechanism; the proposed method has a higher detection accuracy and can perform a more detailed segmentation compared to the baseline. The proposed method was improved based on

the advanced and representative Mask Scoring R-CNN algorithm. By introducing an attention module, using more reasonable feature fusion, and more effective detection branches, the effectiveness and accuracy of the detection of small airport targets were improved significantly. We carried out a series of ablation and comparative experiments on a small-target dataset. The major conclusions are as follows.

(1) Considering that there are currently almost no datasets suitable for small-target detection, we produced a small-target dataset for aprons that meets the definition of small targets from both relative and absolute scales.
(2) Considering that the targets are too small and their features are difficult to extract, an attention module was used to effectively focus on the features and thereby improve the detection accuracy. Furthermore, using a feature fusion module helped achieve a more effective multiscale feature fusion of small targets and a new classifier head was used for a more efficient detection of single classes of small targets and for reducing model complexity.
(3) Considering the difficulties associated with small-target detection in practical applications, a suitable method is proposed. Experiments show that the proposed method achieved improved detection accuracy compared with the baseline. The accuracy in the scale range of (0, 0.01) was improved by 16%, and in the range of (0.01, 0.02), by 15.15% compared with the Mask Scoring R-CNN.

Although the improved method was initially validated by ablation and comparative experiments with promising performance, there is still much room for further development, which is summarised as follows.

(1) The first task will be to further optimise the model and strengthen the practical application of the model, such as decreasing inference speed by reducing model parameters.
(2) Since the targets of this study are pedestrians on the airport surface, the targets are single; thus, the next step can be to expand the dataset for other small-target detection tasks.

Author Contributions: Conceptualisation, H.L.; methodology, H.L.; software, H.L.; validation, H.L.; formal analysis, H.L.; writing—original draft preparation, H.L.; writing—review and editing, H.L., M.D. and A.N.K.; visualisation, H.L. and S.L.; supervision, M.D., S.G. and Y.X.; project administration, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is co-supported by the National Natural Science Foundation of China (No. U2033201 and U2033216). It is also supported by the Opening Project of Civil Aviation Satellite Application Engineering Technology Research Center (RCCASA-2022003) and Nanjing University of Aeronautics and Astronautics Innovation Program Project (xcxjh20220736).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the grants for this manuscript are still in the research phase, and some research data or key codes are currently limited to disclosure within the project team. However, some data involved in this study can be provided. If necessary, you can contact Hao, Liu via email (liuhao_lbj@nuaa.edu.cn) to obtain the Baidu Netdisk (Baidu Cloud) URL link and then download the files you need. Once the link is gone, you can contact the authors of this article to obtain the latest link.

Acknowledgments: We are very grateful for all fundings' assistance. Meanwhile, we appreciate Shuai Li, Yubin Xu, Shuli Gong, and Abdul Nasser Kasule.

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

References

1. Li, X.; Qian, Y.; Chen, H.; Zheng, L.; Wang, Q.; Shang, J. An Unsupervised Learning Approach for Analyzing Unsafe Pilot Operations Based on Flight Data. *Appl. Sci.* **2022**, *12*, 12789. [CrossRef]
2. Izdebski, M.; Gołda, P.; Zawisza, T. The Use of Simulation Tools to Minimize the Risk of Dangerous Events on the Airport Apron. *Adv. Solut. Pract. Appl. Road Traffic Eng.* **2023**, *91*, 107.
3. Lyu, Z.; Zhang, Y. A novel temporal moment retrieval model for apron surveillance video. *Comput. Electr. Eng.* **2023**, *107*, 108616. [CrossRef]
4. Meng, D.; Yuan, D.; Li, W.; Yi, X.; Yun, C. Individual Surveillance around Parked Aircraft at Nighttime: Thermal Infrared Vision-based Human Action Recognition. *IEEE Trans. Syst. Man Cybern. Syst.* **2023**, *53*, 1084–1094.
5. Lu, Z.; Huang, Z.; Song, Q.; Ni, H.; Bai, K. Infrared small target detection based on joint local contrast measures. *Optik* **2023**, *273*, 170437. [CrossRef]
6. Basheer, I.; Zaghdoud, R.; Ahmed, S.; Sendi, R.; Alsharif, S.; Alabdulkarim, J.; Krishnasamy, G. A real-time computer vision based approach to detection and classification of traffic incidents. *Big Data Cogn. Comput.* **2023**, *7*, 22. [CrossRef]
7. Kožović, V.; Đurđević, Ž.; Dinulović, R.; Milić, S.; Rašuo, P. Air traffic modernization and control: ADS-B system implementation update 2022: A review. *FME Trans.* **2023**, *51*, 117–130. [CrossRef]
8. Habibi, J.; Amrhar, A.; Gagné, M.; Landry, R.J. Security Establishment in ADS-B by Format-Preserving Encryption and Blockchain Schemes. *Appl. Sci.* **2023**, *13*, 3105. [CrossRef]
9. Zhang, M.; Zhao, D.; Sheng, C.; Liu, Z.; Cai, W. Long-Strip Target Detection and Tracking with Autonomous Surface Vehicle. *J. Mar. Sci. Eng.* **2023**, *11*, 106. [CrossRef]
10. Zhou, J.; Bai, X.; Zhang, Q. Relevancy between Objects Based on Common Sense for Semantic Segmentation. *Appl. Sci.* **2022**, *12*, 12711. [CrossRef]
11. Slama, B.; Abdo, K.; Vignaud, E.; Simonin, A.; Lohan, S.; Obaid, S.; Ellejmi, M. Use of 5G and mmWave radar for positioning, sensing, and line-of-sight detection in airport areas. In Proceedings of the SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
12. Li, W.; Zhang, L.; Wu, C.; Cui, Z.; Niu, C. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1999–2015. [CrossRef]
13. Li, C.; Zhen, T.; Li, Z. Image classification of pests with residual neural network based on transfer learning. *Appl. Sci.* **2022**, *12*, 4356. [CrossRef]
14. Ahmad, J.; QasMarrogy, A. Modeling of an Airport Traffic Control (ATC) Radars Using Mathcad. In Proceedings of the 4th International Conference on Communication Engineering and Computer Science, Coimbatore, India, 15–16 September 2022; p. 44.
15. Kim, C.; Lee, Y.; Park, J.I.; Lee, J. Diminishing unwanted objects based on object detection using deep learning and image inpainting. In Proceedings of the International Workshop on Advanced Image Technology, Chiang Mai, Thailand, 7–9 January 2018; pp. 1–3.
16. 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.
17. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
18. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision–ECCV, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
22. Kaiming, H.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
24. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
25. Gu, J.; Wang, Z.; Kuen, J. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354377. [CrossRef]
26. Zhang, L.G.; Wang, L.; Jin, M.; Geng, X.S.; Shen, Q. Small targets detection in remote sensing images based on attention mechanism and multi-scale feature fusion. *Int. J. Remote Sens.* **2022**, *43*, 3280–3297. [CrossRef]
27. Luo, H.; Wang, P.; Chen, H.; Kowelo, V.P. Small Object Detection Network Based on Feature Information Enhancement. *Comput. Intell. Neurosci.* **2022**, *2022*, 6394823. [CrossRef] [PubMed]
28. Peng, C.; Zhu, M.; Ren, H.; Emam, M. Small Object Detection Method Based on Weighted Feature Fusion and CSMA Attention Module. *Electronics* **2022**, *11*, 2546. [CrossRef]

29. Bosquet, B.; Cores, D.; Seidenari, L.; Brea, V.M.; Mucientes, M.; Del Bimbo, A. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognit.* **2023**, *133*, 108998. [CrossRef]
30. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for targets detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.
31. Wang, X.; Zhang, R.; Kong, T. SOLOv2: Dynamic, faster and stronger. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
32. Laradji, I.H.; Rostamzadeh, N.; Pinheiro, P.O.; Vázquez, D.; Schmidt, M. Instance segmentation with point supervision. *arXiv* **2019**, arXiv:1906.06392.
33. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: Segmenting targets by locations. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 649–665.
34. Bao, Y.; Song, K.; Liu, J.; Wang, Y.; Yan, Y.; Yu, H.; Li, X. Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5011111. [CrossRef]
35. Tabernik, D.; Šela, S.; Skvarč, J.; Skočaj, D. Segmentation-based deep-learning approach for surface-defect detection. *J. Intell. Manuf.* **2020**, *31*, 759–776. [CrossRef]
36. Aslam, Y.; Santhi, N.; Ramasamy, N.; Ramar, K. Localization and segmentation of metal cracks using deep learning. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 4205–4213. [CrossRef]
37. Han, H.; Gao, C.; Zhao, Y.; Liao, S.; Tang, L.; Li, X. Polycrystalline silicon wafer defect segmentation based on deep convolutional neural networks. *Pattern Recognit. Lett.* **2020**, *130*, 234–241. [CrossRef]
38. Dong, Y.; Wang, J.; Wang, Z.; Zhang, X.; Gao, Y.; Sui, Q.; Jiang, P. A deep-learning-based multiple defect detection method for tunnel lining damages. *IEEE Access* **2019**, *7*, 182643–182657. [CrossRef]
39. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
40. Chen, L.C.; Hermans, A.; Papandreou, G.; Schroff, F.; Wang, P.; Adam, H. Masklab: Instance segmentation by refining targets detection with semantic and direction features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4013–4022.
41. 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.
42. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
43. Liong, S.T.; Gan, Y.S.; Huang, Y.C.; Yuan, C.A.; Chang, H.C. Automatic defect segmentation on leather with deep learning. *arXiv* **2019**, arXiv:1903.12139.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.