applied sciences

MDPI

Article

An Anchor-Free Network for Increasing Attention to Small Objects in High Resolution Remote Sensing Images

Huaping Zhou, Wei Guo * and Qi Zhao

School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China
* Correspondence: 2020201025@aust.edu.cn; Tel.: +86-134-7006-6500

Abstract: Aimed at the problems of small object detection in high resolution remote sensing images, such as difficult detection, diverse scales, and dense distribution, this study proposes a new method, DCE_YOLOX, which is more focused on small objects. The method uses depthwise separable deconvolution for upsampling, which can effectively recover lost feature information and combines dilated convolution and CoTNet to extract local contextual features, which can make full use of the hidden semantic information. At the same time, EcaNet is added to the enhanced feature extraction network of the baseline model to make the model more focused on information-rich features; secondly, the network input resolution is optimized, which can avoid the impact of image scaling to a certain extent and improve the accuracy of small object detection. Finally, CSL is used to calculate the angular loss to achieve the rotated object detection of remote sensing images. The proposed method in this study achieves 83.9% accuracy and 76.7% accuracy for horizontal object detection and rotationally invariant object detection, respectively, in the DOTA remote sensing dataset; it even achieves 96% accuracy for rotationally invariant object detection in the HRSC2016 dataset. It can be concluded that our algorithm has a better focus on small objects, while it has an equally good focus on other objects and is well suited for applications in remote sensing, and it has certain reference significance for realizing the detection of small objects in remote sensing images.

Keywords: YOLOX; depthwise separable deconvolution; dilated convolution; small object detection; CSL

check for updates

Citation: Zhou, H.; Guo, W.; Zhao, Q. An Anchor-Free Network for Increasing Attention to Small Objects in High Resolution Remote Sensing Images. *Appl. Sci.* **2023**, *13*, 2073. https://doi.org/10.3390/app13042073

Academic Editor: Sungho Kim

Received: 12 December 2022
Revised: 31 January 2023
Accepted: 31 January 2023
Published: 5 February 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

The content of satellite images consists of buildings, roads, vehicles, ships, etc. Remotely sensed imagery may combine high spatial resolution with mediocre temporal and usually only very poor spectral resolution, or encompass macroscopic, objective, comprehensive, real-time, movement, and fast elements [1], which provides new means for earth resource investigation and development [2], land improvement [3], environmental monitoring [4], and global research [5].

On the other hand, with the in-depth research of AI theory and DL technology, object detection technology has also made great progress [6]. Object detection is the use of image processing, deep learning, and other techniques to locate the object of interest from an image or video, determine whether the input image contains the object by object classification, find out the location of the target object with object localization, and frame the object, and its task is to lock the object in the image, locate the object location, and determine the object class [7]. With the intensive application of deep convolutional neural networks [8], they have become a powerful tool for object detection. As the key point of image and video understanding, object detection is the primary difficulty in solving more extended tasks, such as image segmentation [9], object tracking [10], image description [11], event detection [12], and scene understanding [13].

However, the focus of object detection differs for remote sensing images and natural images [14]. Due to the differences in imaging platforms and imaging methods, remote

sensing images have complex and diverse backgrounds, large differences in object scales, and arbitrary directions, which leads to poor detection results [15].

At the same time, the resolution of many objects in remote sensing images is quite low, and the detection of small objects has long been one of the key and most difficult points in object detection [16], because small objects have the characteristics of less image coverage area, insufficient resolution, lack of accuracy in location, and inadequate feature expression. Specifically, they are more difficult to detect compared to more extended objects.

Currently, the mainstream deep learning object detection algorithms are mainly divided into two categories, whose main distinction is in the presence or absence of anchors [17].

For the anchor-based approach, however, much is dependent on the number of different samples and the hyperparameters of the anchor points in terms of detection performance, and the non-maximum suppression (NMS) algorithm [18] has been introduced into the detection process to eliminate duplicate object frames, which makes the complexity and computation of the algorithm increase accordingly, limiting its popularity and any improvement in detection speed.

To improve the flexibility of the detector, anchor-free methods such as CornerNet [19], FCOS [20], and CenterNet [21] have emerged and received wide attention. At present, anchorless methods have been partly studied in remote sensing image detection. Li Jie et al. [22] proposed a structure for the feature sharing of parallel layers for aircraft objects, using an attention mechanism, and it maximizes the object characterization capability while maintaining the model complexity and detection speed. Wei Wei et al. [23] discarded the common standard convolution while using depth-separable convolution in the CenterNet residual module, which effectively reduces the network computation and redundancy. They also use an attention mechanism to suppress useless information and provide the network with stronger detection performance. Zhe Zheng et al. [24] used multi-scale attention FPNs in their own networks to suppress noise while enhancing the multiplexing of effective features. The GVR mechanism is used to rotate the detection frame to fit the object more closely. Liu Gaotian et al. [25] optimized the feature extraction network of RFBNet by abandoning the ordinary regular convolution and using self-correcting convolution, which can extend the perceptual field. Using the idea of a multi-scale and dense prediction module to make shallow information richer and integrate contextual information, Lim et al. [26] added an attention mechanism to the network and combined it with contextual features to make the network model more focused on objects with low resolution and deliver more targeted objects.

Inspired by the above research methods, in order to improve the performance of the network model for small object detection in high resolution remote sensing images, and to solve the problems of its diverse scale and dense distribution, this paper proposes a new solution called "Enhanced YOLOX Network Model with Different Convolutions" (DCE_YOLOX), using YOLOX as the baseline model, which uses depth-separable deconvolution for upsampling, combines dilated convolution and CoTNet to extract local contextual features, adds EcaNet to the enhanced feature extraction network of the original network model, and, based on the characteristics of remote sensing images, i.e., the input of the network model, the size is optimized. Through these methods, the network model is made to focus more on information-rich object features and on small target objects with stronger feature extraction capability and higher small object detection accuracy. Finally, the angular loss is calculated by CSL to realize the detection of rotated objects for remotely sensed imagery. The main contributions of this article are as follows:

1. We propose an upsampling method using depth-separable deconvolution. With this special way of expanding the feature map, the pixel values around the central pixel can be diversified when the resolution is expanded, so that the obtained feature map is closer to the original map.
2. We replace the conventional convolution in Contextual Transformer Networks with dilated convolution, which can make better use of local context information to extract

the feature information of small objects in remote sensing images and use EcaNet to improve the feature extraction ability of the network while maintaining the parameters of the original network model.
3. We studied the impact of network input size on the performance of the network model and optimized the network input size. At the same time, we used circular smooth tags to achieve more timely remote sensing image rotation object detection.
4. By comparing the performance of the related algorithms of horizontal object detection and rotationally invariant object detection, the proposed method has the best performance in remote sensing image object detection.

The remaining chapters of this paper are organized as follows: Section 2 introduces the selection of baseline models, Section 3 describes the methods we proposed and used, and Section 4 introduces and analyzes the experimental environment, data, and results in detail, including horizontal object detection and selected object detection. Section 5 summarizes the research contents of this paper.

2. Selection of Baseline Model

YOLOX [27] is an anchor-free version of YOLO; the first version was officially released in July last year, and its performance exceeds that of all previous YOLO series. It draws on many advantages of YOLO series networks; free from a priori frame constraints, its detection speed is faster, detection accuracy is higher, and end-to-end deployment is more flexible.

The following are the four most important components of the network model:

1. Input

YOLOX uses two data enhancement technologies, Mosaic and Mixup, which were used in previous YOLO versions.

2. Backbone

YOLOX uses the Darket-53 network. It has a total of 53 convolutional layers, and the last layer is the full connection layer. The remaining 52 convolutions are used as the main network.

3. Neck

The structure of FPN is used for fusion. FPN allows the features of the different layers of the pyramid to be displayed, such as high semantic features and high-resolution features. the presence of FPN allows the network model to focus on small objects, further improving the performance of the algorithm.

4. Prediction

YOLOX uses decoupled head, anchor-free detector, label assignment strategy SimOTA (Simplified Optimal Transport Assignment), and loss calculation.

For YOLOX, the basic structure of all four models is the same; for all models, X-s, X-m, X-l, or X-x, the network model part is the same. The only difference is that the depth and width of the models are set differently, controlled by two parameters: depth and width. The width and depth of the four models are 0.33×0.50, 0.67×0.75, 1.0×1.0, and 1.33×1.25, respectively, from small to large.

YOLOX applies some of the industry's latest and most advanced technologies to the development of detectors, yielding the best results available on most models. Compared to the previous YOLO series, YOLOX offers faster detection, higher detection accuracy, and more flexibility in end-to-end deployment. The superior performance and flexibility were decisive factors for choosing YOLOX as the base network in this paper.

YOLOX has been updated and improved several times within six months after its release. In version 0.1.1, image caching was supported for faster training, preprocessing was optimized for faster training, and the old distortion enhancement was replaced with the new HSV aug for faster training and better performance. In this work, the baseline of this paper is YOLOX version 0.2.0, updated in January 2022.

3. Materials and Methods

In this section, we first propose a new upsampling method using deep separable deconvolution, and then we propose a new feature extraction method combining dilated convolution and CoTNet. Next, we add EcaNet to the enhanced feature extraction network of the original network model. Finally, we optimize the input size of the network model according to the characteristics of remote sensing images and use CSL to calculate the angle loss. The rotation object detection of remote sensing images is realized.

3.1. New Upsampling Method

When using the deep convolution neural network to classify and predict each pixel in an image, it is necessary to keep the resolution of the output predicted image consistent with the input image, so it is necessary to upsample the feature map. The YOLO network uses nearest neighbor interpolation to upsample the feature maps, which leads to ambiguous region semantics, insensitivity to details, and easy loss of objects. Although deconvolution cannot restore the feature map, it has similar effects and can effectively recover a small part of the lost information to the maximum. However, it brings higher parameter quantity to the network model than upsampling using nearest neighbor interpolation.

Howard et al. [28] first introduced deep separable convolution, which decomposes the standard convolution into deep convolution and point by point convolution, reducing the number of parameters. Because deep separable convolution can extract the channel information more carefully and deeply, the network accuracy is further improved.

In this paper, we propose an upsampling method using depth-separable deconvolution. It can reduce the number of model parameters and slightly increase the model accuracy through deep separable deconvolution while upsampling. With this special way of expanding the feature map, the pixel values around the central pixel can be diversified when the resolution is expanded, so that the obtained feature map is closer to the original map, thereby improving the accuracy of the network model.

In order to further reduce the improvement of parameter quantity caused by deconvolution, we apply group convolution to our deep separable deconvolution and adjust the convolution kernel size of deconvolution. Finally, our method adjusts the size of the feature map and improves the accuracy of the network model on the premise of adding a few parameters. Figure 1 shows the principal diagram of deep separable deconvolution. In the model, we adjust the number of corresponding deconvolution groups according to the number of channels in the network.

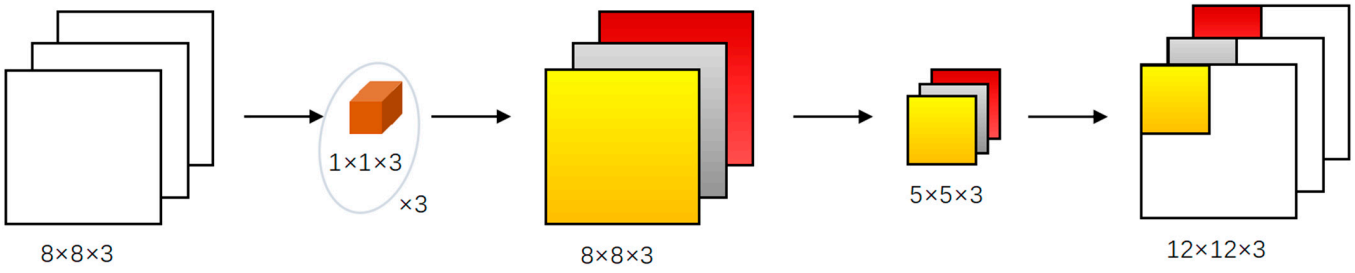


Figure 1. Schematic diagram of deep separable deconvolution.

3.2. New Feature Extraction Method

Due to the limitations of conventional convolution operation, the method based on the fusion of CNN and transformer has attracted wide attention. It solves the problems of poor interaction between CNN and global and contextual semantic information, and the loss of features of transformer. However, most of these methods ignore the rich context information between adjacent keys. In view of these problems, CoTNet integrates local context information and global context information on the basis of traditional self-attention learning to better enhance visual expression ability. Compared with the commonly used conv, it has better feature extraction ability.

Dilated convolution inserts 0 between each parameter of standard convolution. On the basis of guaranteeing constant standard convolution parameters, the perception field is expanded so that each convolution kernel can obtain a larger range of information, thus avoiding the loss of spatial location information caused by the pooled layer. The convolution of holes with smaller expansion rates can extract small object information from the image, which is helpful for accurate classification of small objects and enhances the model's ability to detect small objects.

To solve the problems of the large resolution of existing remote sensing images, small convoluted sensing fields in the network, and insufficient utilization of local context information, we propose a new feature extraction method, namely DCoTNet. It replaces the conventional convolution in CoTNet (Contextual Transformer Networks) [29] with dilated convolution, which can better extract the feature information of small objects in remote sensing images and improve the detection performance of the network. The formula for the new method is as follows:

$$Y = K + XE_V * [K\ ,\ X]C_rC_o \tag{1}$$

where Y is the output result and X is the characteristic channel of the input, K is the local context information obtained by the dilated convolution, E_V is the embedded matrix, and C_r and C_o are two 1×1 convolutions with and without relu activation function, respectively. Figure 2 shows the schematic diagram of DCoTNet. The width, height, and channel number of input and output have not changed.

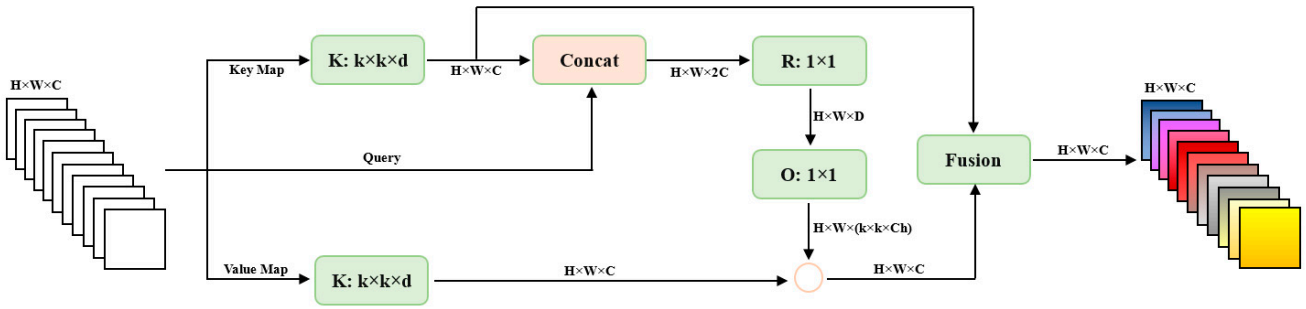


Figure 2. The detailed structures of DCoTNet.

3.3. Introduction of EcaNet

Attention mechanisms have been proven to be an effective method to improve deep convolutional neural networks; channel attention, in particular, plays an extremely important role. Attention mechanisms such as SENet [30] and CBAM [31] have obtained effective performance through complex structural design.

Based on YOLOX, this paper uses the effective channel attention module of the EcaNet (effective channel attention for deep convolutional neural networks) [32] deep convolution neural network. Different from the above attention mechanism, the purpose of ECA is to learn effective attention channels while reducing model complexity.

In channel attention, there is local periodicity of channel features, and dimensionality reduction can negatively affect the network's ability to learn inter-channel relationships. The changes in channel dimension and the interactions between channels can affect the performance of channel attention to some extent. Keeping the channel dimension constant supports network learning. The effective attentional cross-channel interaction can reduce network complexity without performance loss. ECA reduces the complexity of the algorithm and improves the efficiency of learning attention by adding local cross-channel interaction and channel sharing parameters. The channel weights are calculated as follows:

$$\omega_i = \sigma(\sum_{j=1}^{k} \alpha_j y_i^j)\ ,\ y_i^j \in \Omega_i^k \tag{2}$$

where Ω_i^k is the k-domain channel of y_i, y_i is the feature representation of channel i after global average pooling, α is the shared parameter, ω_i is the weight of channel i, and σ is the Linear rectification function.

The k value as a key parameter can be adjusted to determine the range of interaction between channels by adjusting its own size, and the range of interaction increases with the increase in channel dimension. Assuming that the interaction range, k, is proportional to the channel dimension, C, the relationship between them can be extended to a nonlinear one, as in the formula:

$$C = \varnothing(k) = 2^{(r \times k - b)} \tag{3}$$

The k value denotes the convolution kernel size, which can be calculated adaptively as:

$$k = \left| \frac{log_2 C + b}{\gamma} \right|_{odd}, \ \gamma = 2, \ b = 1 \tag{4}$$

In the formula, C is the number of channels, $|\cdot|_{odd}$ means to take the nearest odd number, γ and b are the hyperparameters, and the best results are obtained by experiments with $\gamma = 2$, $b = 1$.

The ECA operation is expressed as:

$$\mathrm{Eca}(x) = \sigma\left(Conv_{1d}^{1\times1}(\mathrm{GAP}(x))\right) \times x \tag{5}$$

In the formula, GAP denotes the global average pooling operation, and x is the input tensor.

Figure 3 shows the general illustration of the EcaNet method, where the EcaNet module is set between the backbone network and the enhanced feature extraction network of YOLOX after experimental validation.

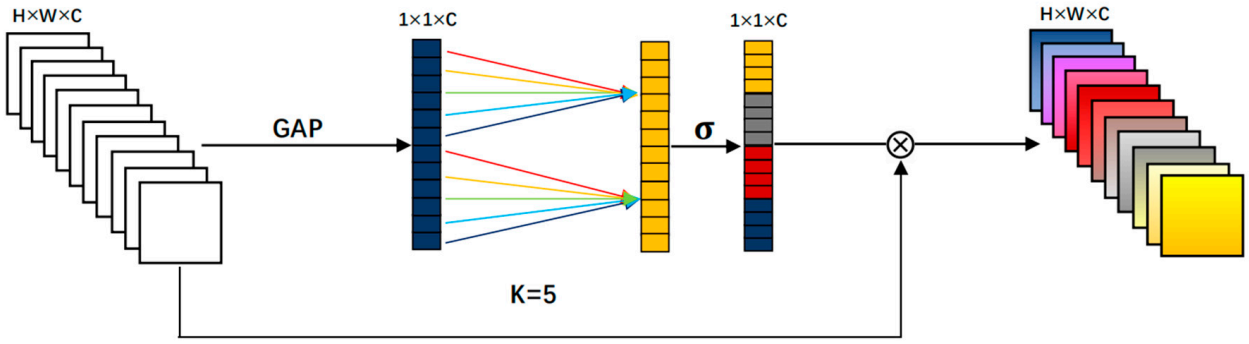


Figure 3. Effective channel attention net.

3.4. Optimized Network Input Size

Most object detection methods are first pre-trained on ImageNet [33], after which the weight parameters obtained from the pre-training are used in formal experiments. Starting from AlexNet [34], most classifiers operate on input images smaller than 256 × 256. YOLOv1 [35] pre-trained the main part of the model on the ImageNet classification dataset using a network input size of 224 × 224, and later increased the network input size to 448 × 448 for detecting objects. However, this approach may make it difficult for the network model to quickly adapt to the high size network input. In contrast, YOLOv2 [36] used an input network size of 416 × 416. The network input size of 224 × 224 is used in pre-training, and then the model is fine-tuned with 10 epochs using an input size of 448 × 448 in order for the network model to work better on the higher network input size. In YOLOv3 [37] and YOLOv4 [38], network input sizes of 416 or 608 are mostly used for better detection of multiple small objects. YOLOv5 most often uses a network input size of 640, using three different scales of feature maps, namely 20, 40, and 80, to detect more fine-grained features of the object.

Due to the visual details of high-resolution images, the high image input size model performs better for object detection with low resolution. However, the low image input size model has better results for large object detection than the high-resolution model, because the backbone network captures more information in the whole image. Due to the design of the currently used FPN structure, the feature map will handle objects of corresponding sizes. Experimental results by Qi et al. [39] show that small objects perform better with large input size, while large objects perform better with small input size. This phenomenon also shows, to some extent, that the divide-and-conquer strategy of the FPN structure used to assign superparameters is not an optimal solution.

In the YOLO network, the process of downsampling goes through five iterations, so the network input size is generally set to the 5th power of 2, which is a multiple of 32. Figure 4 shows a DCE_YOLOX network model with an input size of 1024 × 1024.

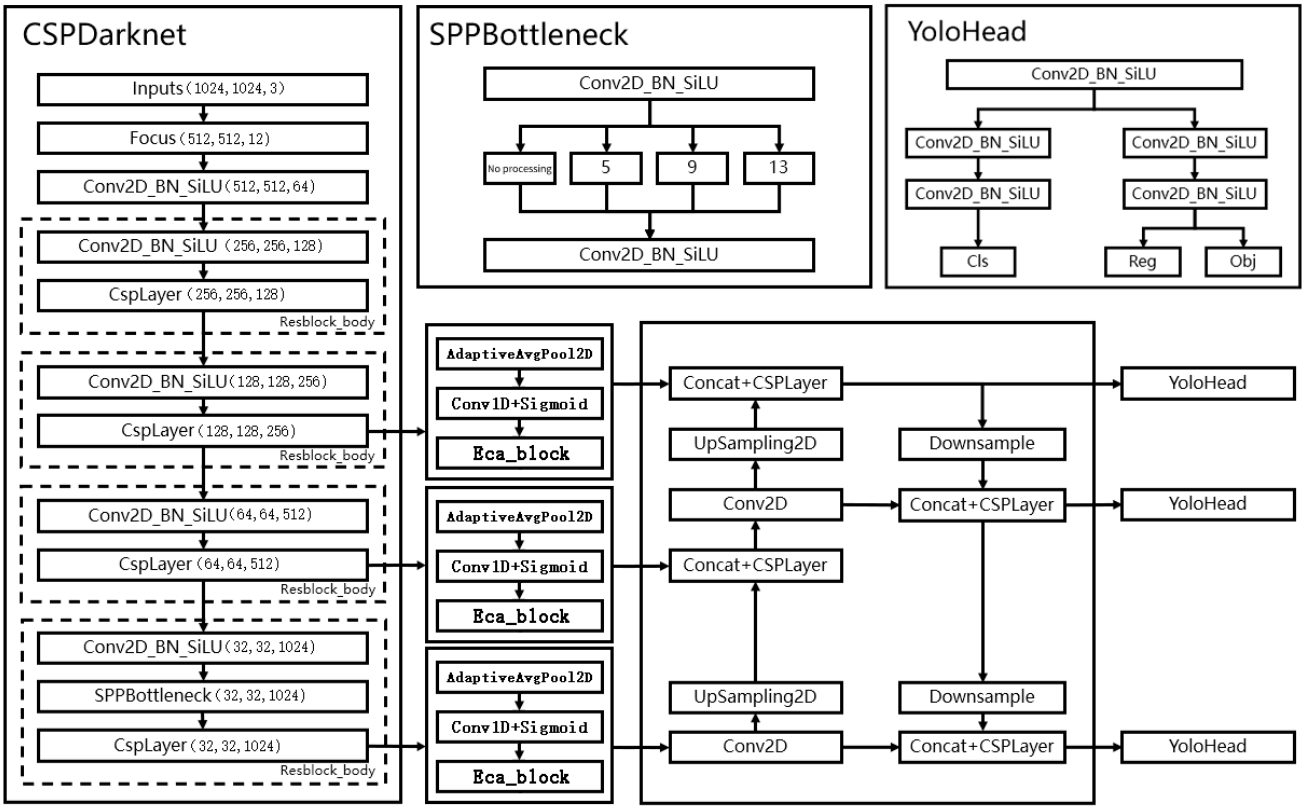


Figure 4. DCE_YOLOX network structure.

3.5. Rotationally Invariant Object Detection

Currently, there are three kinds of object detection bounding boxes: horizontal bounding box, rotating bounding box, and custom bounding box. One of the characteristics of remote sensing image object detection is that the direction of the object to be detected is random and diverse, and the boundary box labeling method adopted by the object detection method should be changed according to the shape characteristics of the detected object itself. Obviously, in view of the fact that the direction of the object is not fixed in remote sensing image object detection, rotating the bounding box is a better choice. In this paper, because the remote sensing image dataset used has undergone data preprocessing, the image is clipped into a rectangle, and the rotation angle of the target is defined as the angle where the x-axis clockwise rotation meets the longest edge of the detection frame.

Rotationally invariant object detection needs to introduce new angle loss. Common definition methods of arbitrary rotating frame include the x, y, w, h, and angle θ Opencv definition method and the long side definition method of five parameters, including four vertex coordinates sorted counterclockwise and an ordered quadrilateral definition method

with eight parameters. Due to the periodicity of the angle, boundary discontinuity will occur when calculating the angle loss, resulting in the sudden increase in the loss value. The angle loss was previously calculated using mostly One-hot, but we found that the angle loss of -90° and 89° differed significantly from each other, even though they were actually only 1° apart.

In addition to the common horizontal boundary box object detection, this paper also carries out rotation boundary box object detection, adding the angle loss. The loss is mainly divided into four parts: regression loss, confidence loss, classification loss, and angle loss. CSL (circular smooth label) [40] is introduced in the calculation of angle loss, which solves the regression problem of angle in the form of classification. By dividing the angle, the prediction result is limited, and the boundary problem is eliminated. The circular smooth label obtains the angle prediction by classification without affecting the boundary conditions. The cyclic label coding method used by the circular smooth label has periodicity, and the assigned smooth label value has a certain tolerance. The expression of CSL is as follows:

$$\mathrm{CSL}(x) = \begin{cases} g(x), & \theta - r < x < \theta + r \\ 0, & otherwise \end{cases} \tag{6}$$

Among them, $g(x)$ is a window function, and its four main characteristics are periodicity: $g(x) = g(x + kT)$, $k \in N$, $T = 180/\mathrm{w}$; symmetry: $0 \le g(\theta + \varepsilon) = g(\theta - \varepsilon) \le 1$, $|\varepsilon| < r$; maximum value: $g(\theta) = 1$; and monotonicity: $0 \le g(\theta \pm \varepsilon) \le \mathrm{g}(\theta \pm \varsigma) \le 1$, $|\varsigma| < |\varepsilon| < r$. The r represents the radius of $g(x)$ and θ represents the angle of the current bounding box. Due to the setting of the window function, $g(x)$, the network model can calculate the angular distance between the predicted tag and the real tag, so that the loss value of the predicted value decreases with the improvement of the prediction result. The problem of angular periodicity is solved by the periodicity of $g(x)$. The CSL is shown in Figure 5.

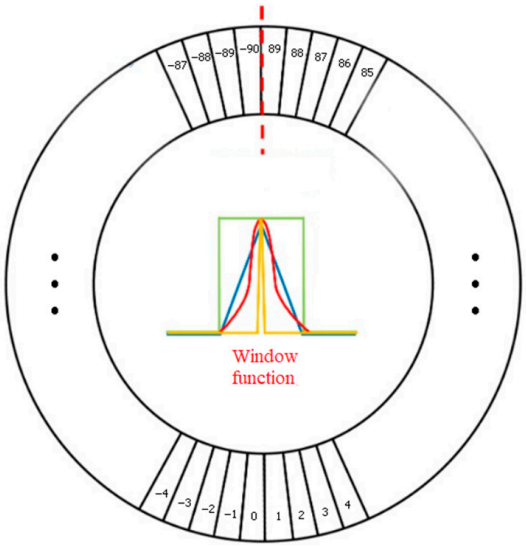


Figure 5. Circular smooth label.

4. Experiments, Results, and Discussion

4.1. Dataset and Environment

We demonstrate the performance of the improved algorithm through comparative experiments on the publicly available datasets DOTA [41] and Hrsc2016 [42]. The DOTA dataset contains 2806 images of different scales, with resolutions as small as 800×800 and as large as 4000×4000. It contains 15 categories of data, with a total of 188,282 instances.

The Hrsc2016 dataset is a remote sensing image dataset for ships. It mainly includes various ships on the sea or on the shore. It was released by Northwest Polytechnic University in 2016. There are 1061 remote sensing images with different resolutions from

300 × 300 to 1500 × 900, including a training set with 436 instances and a test set with 181 instances, and 444 test sets with 2976 instances.

We resized the image according to the image pyramid and cropped the DOTA dataset into 21046 pcs sub-images of 1024 × 1024 size. Similarly, we trimmed the Hrsc2016 dataset to 768 × 768 size sub images. During the experiment, we used the DOTA dataset for horizontal object detection and the DOTA dataset and Hrsc2016 dataset for rotationally invariant object detection. Table 1 shows the relevant tools and the configurations used during the experiments.

Table 1. Hardware and software.

Parameters	Configuration
Operating System	Ubuntu 20.04.2 LTS
CPU	AMD Ryzen 5 3600
GPU	GeForce RTX 3060 Ti
Languages	Python
Platform	CUDA11.1,CuDNN8.0
Framework	Pytorch 1.7.0,Torchvision 0.8.1

4.2. Evaluation Indicators

We used average precision (AP) and mean accuracy (mAP), which are more common in object detection algorithms, to measure the effectiveness of this algorithm. The AP is related to the precision, which is the ratio of positive samples correctly predicted in the prediction dataset to the positive samples predicted by the model, and the recall, which is the ratio of positive samples correctly predicted in the prediction dataset to the actual positive samples. The above measures were calculated as follows:

$$\mathrm{AP} = \int_0^1 Pdr \tag{7}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{8}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} = \frac{TP}{all\ detctions} \tag{9}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} = \frac{TP}{all\ ground\ truths} \tag{10}$$

The value of mAP is obtained by averaging the APs of all categories; N indicates the total number of detected categories, and N = 15 in this experiment. The value of mAP is proportional to the detection effect of the algorithm and the recognition accuracy. *TP*, *FP*, and *FN* indicate the proportion of correct, incorrect, and missing detection frames, respectively.

4.3. Experimental Results

4.3.1. Horizontal Object Detection

In this experiment, four different models, YOLOX-s, YOLOX-m, YOLOX-l, and YOLOX-x, were used in the YOLOX model cluster, as well as the improved models in this paper, DCE_YOLOX-s, DCE_YOLOX-m, DCE_YOLOX-l, and DCE_YOLOX-x. The training batch size was set to 32, and the performance metrics were tested after every 10 epochs of training were completed, and a total of 100 periods of baseline and improved models were trained. Table 2 shows the comparison experiments of the four versions, and the best results are highlighted in bold.

Table 2. This Detailed results of mAP for each model category.

mAP \ Models	YOLOX-s	YOLOX-m	YOLOX-l	YOLOX-x	DCE_ YOLOX-s	DCE_ YOLOX-m	DCE_ YOLOX-l	DCE_ YOLOX-x
small-vehicle	0.5320	0.4412	0.6301	0.5458	0.7402	0.7319	**0.7853**	0.7518
large-vehicle	0.8312	0.8186	0.8655	0.8081	0.8732	0.8781	**0.8903**	0.8729
plane	0.8953	0.9042	0.9013	0.9005	0.9019	**0.9049**	0.9024	0.9041
storage-tank	0.7367	0.7883	0.7698	0.7852	0.8351	**0.8715**	0.8673	0.8686
ship	0.8442	0.8929	0.8597	0.8558	0.8964	**0.9001**	0.8865	0.8975
harbor	0.8103	0.8647	0.8418	0.8251	0.8645	0.8803	0.8799	**0.8816**
ground-track-field	0.7224	0.7864	0.7769	0.7747	0.7299	0.7808	0.7801	**0.8063**
soccer-ball-field	0.7234	0.7858	0.7621	0.7646	0.8112	0.8348	0.8125	**0.8389**
tennis-court	0.9043	0.9064	0.9048	0.9061	0.9064	0.9075	**0.9076**	0.9069
swimming-pool	0.7115	0.7722	0.7525	0.7199	0.7670	0.7425	**0.8046**	0.8030
baseball-diamond	0.7547	0.8034	0.7573	0.7593	0.7765	0.7902	0.8077	**0.8308**
roundabout	0.7100	0.7371	0.7594	0.7630	0.7203	0.7712	0.7761	**0.8088**
basketball-court	0.7697	0.8823	0.8570	0.8305	0.8741	**0.8996**	0.8897	0.8989
bridge	0.5795	0.6787	0.6380	0.6562	0.6518	0.7212	0.6942	**0.7294**
helicopter	0.7976	**0.8490**	0.8422	0.8316	0.7484	0.8236	0.8432	0.7850
all classes	0.7549	0.7941	0.7946	0.7818	0.8065	0.8292	0.8352	**0.8390**

As can be seen from Table 2, among the eight different YOLOX models, the improved model has some improvement in performance indexes compared with its corresponding initial model, among which the improved X-x has the largest improvement, with a 5.72% improvement in mAP, and the mAPs of X-s, X-m, and X-l have improved by 5.16%, 3.51%, and 4.06%, respectively. Therefore, it can be proved that the improved network model has better detection performance than the original network model. In addition, the cross-sectional comparison of the four improved models shows that DCE_YOLOX-x has the best performance.

We compared the performance and speed of the model before and after improvement, including mAP_0.5, mAP_0.5:0.95, Speed, and FLOPs, where Speed is the sum of average forward time, average NMS time, and average inference time. DCE_YOLOX sacrifices part of the detection speed when improving the detection accuracy, and the detection speed and Floating point operations are inferior to the original model. Among them, the optimal mAP_0.5 and mAP_0.5:0.95 of the improved model are 4.44% and 5.52% higher than the original model, respectively, but the optimal Speed and FLOPs are 12.78 ms and 41.6 G lower compared to the original model. The comparison results are shown in Table 3, and the best results are highlighted in bold.

Table 3. Performance comparison.

Model	mAP 0.5	mAP 0.5:0.95	Speed (ms)	FLOPs (G)
YOLOX-s	75.49	51.71	**14.30**	**26.67**
YOLOX-m	79.41	56.30	26.26	73.55
YOLOX-l	79.46	56.83	39.98	155.37
YOLOX-x	78.18	56.45	68.82	281.59
DCE_YOLOX-s	80.65	57.01	27.08	68.27
DCE_YOLOX-m	82.92	59.71	54.40	188.28
DCE_YOLOX-l	83.52	61.30	91.74	397.76
DCE_YOLOX-x	**83.90**	**62.35**	155.82	720.86

Three sets of representative images from the dataset were selected for testing, and the before-and-after comparison graph shown in Figure 6 was obtained, which further shows the improved performance of the improved network for small object detection.



Figure 6. Comparison chart before and after improvement. The upper part is before improvement, and the lower part is after improvement.

In Figure 6 (left), there are many small objects, and the improved YOLOX before the improvement can only detect larger objects and misses other small objects, while the improved DCE_YOLOX can effectively detect the presence of small objects alone. In Figure 6 (middle), there are many small objects with a dense distribution; the original YOLOX algorithm is poor in detecting dense small objects, but the improved algorithm can detect the dense objects very well. Figure 6 (right) shows small objects with inconspicuous appearance features, which are poorly detected by the algorithm before the improvement; the improved DCE_YOLOX has a small number of missed detections, but the overall detection effect is excellent.

This is due to the fact that DCE_YOLOX introduces the EcaNet effective channel attention module between the backbone network and the enhanced feature extraction network, while boosting the network image input size and adaptively calibrating the weights between channels of different feature layers, effectively fusing deep and shallow features, reducing semantic loss, and still recognizing small objects effectively.

In order to compare the feature extraction ability of DCE_YOLOX and the original model, we visualized the first 32 feature images of both models separately, using the first image in Figure 6 as an example, as shown in Figure 7. We found that the improved model has better discrimination and anti-interference ability for the background, as well as better feature extraction ability for useful targets. This is due to the inclusion of the effective channel attention module in the DCE_YOLOX, which allows the network to make full use of feature information and focus more on targets with distinct features. At the same time, the optimized network input size allows the network to improve the detection performance for small objects.

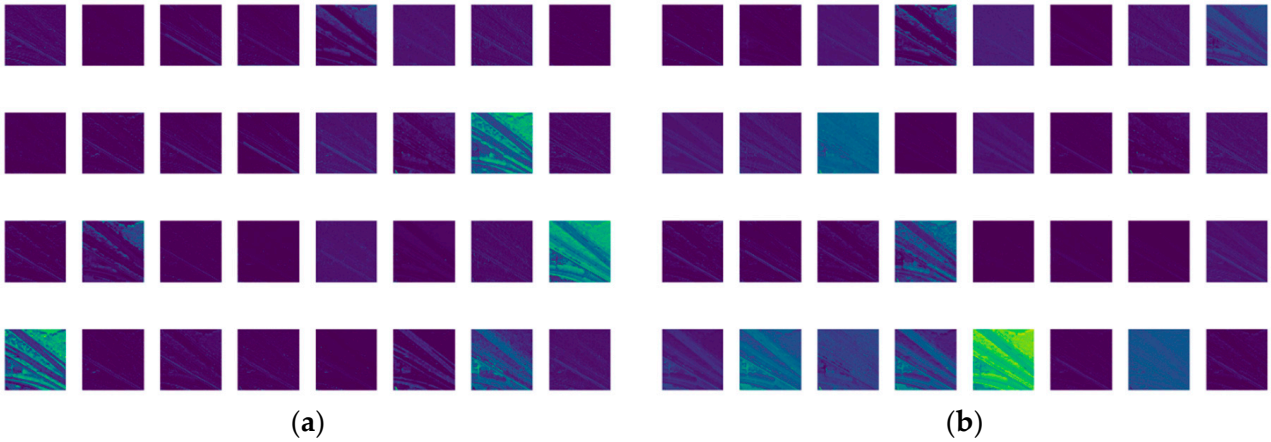


Figure 7. The first 32 feature visualizations: (**a**) is before the improvement, and (**b**) is after the improvement.

To further evaluate the performance of the DCE_YOLOX algorithm, using the DOTA dataset with the same training and testing samples, this paper compares Faster RCNN [43], R-FCN [44], SSD [45], YOLOv2, YOLOv3, and YOLOv5 series, and the comparison results are shown in Table 4; the best results are highlighted in bold.

Table 4. Detection results of different algorithms.

Method	Backbone	mAP
SSD300	VGG16	10.9
SSD512	VGG16	20.8
Faster R-CNN	ResNet101	60.5
R-FCN	ResNet101	47.2
YOLOv2	Darknet19	21.4
YOLOv3	Darknet53	53.7
YOLOv5s	CSPFocus	71.0
YOLOv5m	CSPFocus	73.5
YOLOv5l	CSPFocus	74.4
YOLOv5x	CSPFocus	73.7
DCE_YOLOX-x	CSPFocus	**83.9**

From the table, we can see that DCE_YOLOX has the highest mAP compared to other algorithms, with a detection accuracy of 83.9%. In addition, the DCE_YOLOX algorithm improves 73.0%, 63.1%, 23.4%, 36.7%, 62.5%, 30.2%, and 9.5% in average accuracy mAP, respectively, compared with SSD300, SSD512, Faster R-CNN, R-FCN, YOLOv2, YOLOv3, and the optimal YOLOv5. From these experiments, we can conclude that our algorithm has a better focus on small objects, while it has an equally good focus on other objects and is well suited for applications in remote sensing.

4.3.2. Rotationally Invariant Object Detection

In this experiment, we used YOLOX-s as the baseline model and used the circular smooth label to calculate the angle loss to realize the rotation object detection of remote sensing images. Table 5 shows the detailed comparison between our method and the current common methods on the DOTA dataset. The accuracy of mAP reached 76.41%, even higher than the current SOTA algorithm ReDet [46]. The data in the table show that the precision of DCE-YOLOX is 0.16% higher than that of ReDet on the DOTA dataset. In addition, the memory occupied by our model weight is 68.5 MB, which is lower than all the algorithms in the table.

Table 5. Detection results on DOTA dataset.

Method	Backbone	mAP	Memory
R3Det [47]	ResNet50	70.08	143
ReDet	ResNet50	76.25	125
PolarDet [48]	ResNet50	75.02	150
S2ANet [49]	ResNet50	74.12	148
CFA [50]	ResNet50	73.45	141
Oriented Reppoints [51]	ResNet50	75.97	230.5
Oriented R-CNN [52]	ResNet50	75.87	158
RepVGG-YOLO [53]	RepVGG	74.13	-
DCE-YOLOX	CSPFocus	76.41	68.5

Figure 8 shows the confusion matrix of our method, and we can see the detection accuracy of each category. The best detection results are plane and ship, with an accuracy of 96%, and the worst category is soccer-ball-field, with only 52%. Because the detection of rotated objects is performed, there is a difference in the detection effect of some objects with similar shapes and sizes. Compared to horizontal detection, the angle prediction is added here. In addition, the number of objects in different categories in the dataset varies, which is one of the reasons for the difference in the detection results of different categories of objects. For specific category information and examples, please refer to the relevant documents of the DOTA dataset. It can be seen from the results that our algorithm has a good detection effect on small objects and objects with obvious features in remote sensing images, while the detection of other objects still needs to be improved.

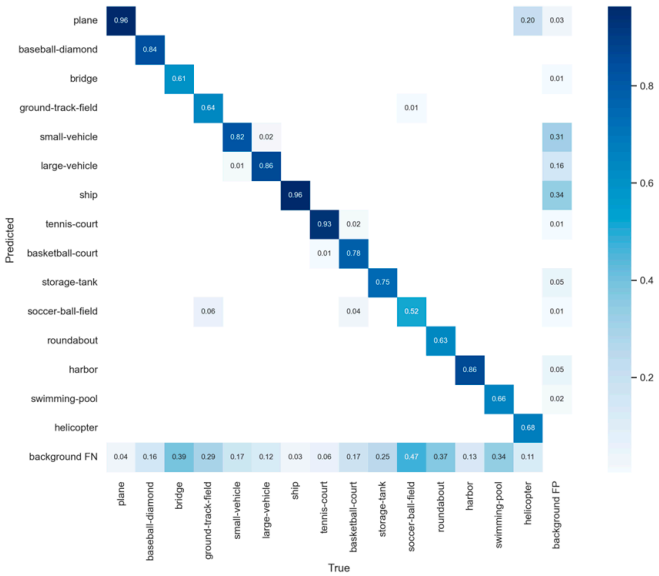


Figure 8. Confusion matrix of DCE_YOLOX.

Table 6 shows the performance comparison results of the HRSC2016 dataset. The detection accuracy of our method is higher than that of other rotationally invariant object detection methods. The detection accuracy reaches 95.9%, and there is almost no false detection or missing detection. At the same time, it has a high detection speed. This shows that our method has a good detection effect in single-category or multi-category detection, which reflects the universality of DCE_YOLOX.

Table 6. Detection results of HRSC2016 dataset.

Method	Backbone	Image Size	mAP	FPS
R2PN [54]	VGG16	-	79.6	-
RRD [55]	VGG16	384 × 384	84.3	-
SARD [56]	ResNet101	800 × 800	85.4	1.5
ROI Transformer [57]	ResNet101	512 × 800	86.2	5.9
R4Det [58]	ResNet50	800 × 800	88.17	8.6
RepVGG-YOLO	RepVGG	-	91.54	22
DCE-YOLOX	CSPFocus	768 × 768	95.9	17.5

Figure 9 shows the visualization results of our method on the DOTA dataset and HRSC2016 dataset. The first three lines are the visualization results of the DOTA dataset, and the fourth line is the visualization results of the HRSC2016 dataset. It can be seen from the figure that our method has a good detection effect, even for dense and small objects.



Figure 9. The visualization results on DOTA dataset and HRSC2016 dataset. The first two lines are the DOTA dataset, and the third line is the HRSC2016 dataset.

4.3.3. Ablation Study

In this experiment, we conducted ablation experiments on our methods on the DOTA dataset to explore the improvement of each method on the network model, mainly including deep separable deconvolution, DCoTNet, EcaNet, optimized network input size, and CSL. Because this paper mainly focuses on the detection of rotating targets, the baseline model for comparison is mainly based on the detection of rotating targets. At the same time, we used Precision, Recall, and mAP as the evaluation indicators of the experiment. The specific experimental results are shown in Table 7.

Table 7. Ablation study on DOTA dataset.

DS-DConv	DCoTNet	EcaNet	O-N-I-S	CSL	Precision	Recall	mAP
				√	0.74904	0.68846	0.72001
			√	√	0.74896	0.70741	0.73304
		√	√	√	0.75729	0.70814	0.73422
	√	√	√	√	0.76333	0.70947	0.76142
√	√	√	√	√	0.77661	0.72814	0.76413

In the experiment, we found that, after the object detection frame is rotated, the mAP is lower than the horizontal object detection. This is due to the fact that performing detection of rotated objects requires additional consideration of the deflection angle of the object, and therefore the detection accuracy is degraded. Although the mAP is reduced, the actual detection results are greatly improved. The optimized network input size has greatly improved the performance of the network model, and the mAP has increased by 1.3%. This is because the large resolution input makes it easier for the network to identify small targets. The corresponding performance improvement will also increase the number of the network parameters and reduce the detection speed. EcaNet has no impact on the scale of the network model and will improve the detection accuracy of the network model in a limited way. It is a very useful plug-and-play network. DCoTNet extracts the semantic and location information of the target, which can improve the detection effect of the network, and the mAP is increased by 2.72%. Compared with ordinary oversampling, deep separable deconvolution can diversify the pixel values around the center pixel when the resolution is expanded, so that the feature map obtained is closer to the original map, thus improving the accuracy of the network model. The size of its convolution kernel has different effects on the results. After many experiments, we balanced the parameter amount and improved the performance. We set it to 4, which can not only provide detection accuracy, but also accelerate the convergence speed of training.

5. Conclusions

In this paper, we optimized YOLOX and proposed an anchor-free object detection network model for detecting objects in remotely sensed images, called DCE_YOLOX. In order to solve the problems of low resolution, complex background, and scale change of the detection object, it uses depth separable deconvolution for upsampling, which can effectively recover the lost feature information while expanding the feature map. Moreover, it combines dilated convolution and CoTNet for feature extraction, makes full use of local context information, and expands the receptive field. It also adds EcaNet to the enhanced feature extraction network of the original network model, uses the feature information through the effective channel attention module, makes the model pay more attention to information-rich features, and improves the detection ability of the network model to small objects. Then, according to the high image resolution characteristics of remote sensing images, the network input size is optimized to improve the detection performance of the model for small objects. Furthermore, CSL was used to calculate the angle loss and realize the rotation object detection of remote sensing images. Finally, experiments were carried out on the multiscale, high image resolution DOTA dataset and the HRSC2016 dataset. Our method pays much more attention to small objects, especially small objects in images. Although the detection speed has decreased, the detection ability exceeds other commonly used algorithms. In the future, we will further optimize the feature representation of the model, enhance the generalization ability of the model, and study different annotation methods, while ensuring faster detection speeds.

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

Funding: This research was funded by Anhui University of Science and Technology, Huaii'nan, China, National Natural Science Foundation of China 61703005, and Anhui Province Key R&D Program of International Science and Technology Cooperation Special Project 202004b11020029.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Anhui University of Science and Technology for its support, Secondly, they thank ITVR-AUST laboratory for its support.

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

References

1. Li, D.; Wang, M.; Jiang, J. China's high-resolution optical remote sensing satellites and their mapping applications. *Geo-Spat. Inf. Sci.* **2021**, *24*, 85–94. [CrossRef]
2. Li, L.J.; Wu, Y. Application of remote-sensing-image fusion to the monitoring of mining induced subsidence. *J. China Univ. Min. Technol.* **2008**, *18*, 531–536. [CrossRef]
3. Ansith, S.; Bini, A.A. Land use classification of high resolution remote sensing images using an encoder based modified GAN architecture. *Displays* **2022**, *74*, 102229.
4. Gong, M.; O'Donnell, R.; Miller, C.; Scott, M.; Simis, S.; Groom, S.; Tyler, A.; Hunter, P.; Spyrakos, E.; Merchant, C.; et al. Adaptive smoothing to identify spatial structure in global lake ecological processes using satellite remote sensing data. *Spat. Stat.* **2022**, *50*, 100615. [CrossRef]
5. Dong, J.; Li, L.; Li, Y.; Yu, Q. Inter-comparisons of mean, trend and interannual variability of global terrestrial gross primary production retrieved from remote sensing approach. *Sci. Total Environ.* **2022**, *822*, 153343. [CrossRef] [PubMed]
6. Wang, Y.; Bashir SM, A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Syst. Appl.* **2022**, *197*, 116793. [CrossRef]
7. 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]
8. Chengji, X.U.; Wang, X.; Yang, Y. Attention-YOLO: YOLO Detection Algorithm That Introduces Attention Mechanism. *Comput. Eng. Appl.* **2019**, *55*, 13–23.
9. Wang, Y.; Gao, L.; Hong, D.; Sha, J.; Liu, L.; Zhang, B.; Rong, X.; Zhang, Y. Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *104*, 102582. [CrossRef]
10. Xuan, S.; Li, S.; Zhao, Z.; Zhou, Z.; Gu, Y. Rotation adaptive correlation filter for moving object tracking in satellite videos. *Neurocomputing* **2021**, *438*, 94–106. [CrossRef]
11. Kumawat, A.; Panda, S. Feature detection and description in remote sensing images using a hybrid feature detector. *Procedia Comput. Sci.* **2018**, *132*, 277–287. [CrossRef]
12. Liu, L.; Li, C.; Sun, X.; Zhao, J. Event alert and detection in smart cities using anomaly information from remote sensing earthquake data. *Comput. Commun.* **2020**, *153*, 397–405. [CrossRef]
13. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulosd, P.T. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 337–350. [CrossRef]
14. 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]
15. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
16. Xiaolin, F.; Fan, H.; Ming, Y.; Tongxin, Z.; Ran, B.; Zenghui, Z.; Zhiyuan, G. Small object detection in remote sensing images based on super-resolution. *Pattern Recognit. Lett.* **2022**, *153*, 107–112. [CrossRef]
17. Tong, K.; Wu, Y. Deep learning-based detection from the perspective of small or tiny objects: A survey. *Image Vis. Comput.* **2022**, *123*, 104471. [CrossRef]
18. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855.
19. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
20. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF INTERNATIONAL Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
21. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
22. Liu, J.; Yang, D.; Hu, F. Multiscale Object Detection in Remote Sensing Images Combined with Multi-Receptive-Field Features and Relation-Connected Attention. *Remote Sens.* **2022**, *14*, 427. [CrossRef]

23. Wei, W.; Ru, Y.; Ye, Z. Improve the remote sensing image target detection of centernet. *Comput. Eng. Appl.* **2021**, *57*, 9.
24. Zheng, Z.; Lei, L.; Sun, H.; Kuang, G. FAGNet: Multi-Scale Object Detection Method in Remote Sensing Images by Combining MAFPN and GVR. *J. Comput.-Aided Des. Comput. Graph.* **2021**, *33*, 883–894. [CrossRef]
25. Shi, P.; Zhao, Z.; Fan, X.; Yan, X.; Yan, W.; Xin, Y. Remote Sensing Image Object Detection Based on Angle Classification. *IEEE Access* **2021**, *9*, 118696–118707. [CrossRef]
26. Lim, J.S.; Astrid, M.; Yoon, H.J.; Lee, S.I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 181–186.
27. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
29. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
31. 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.
32. Wang, Q.; Wu, B.; Zhu, P.; Li, B.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
33. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
35. 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.
36. 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 June 2017; pp. 7263–7271.
37. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
38. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
39. Qi, L.; Kuen, J.; Gu, J.; Lin, Z.; Wang, Y.; Chen, Y.; Li, Y.; Jia, J. Multi-scale aligned distillation for low-resolution detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14443–14453.
40. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 677–694.
41. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
42. Wang, C.; Bai, X.; Wang, S.; Zhou, P. Multiscale visual attention networks for object detection in VHR remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 310–314. [CrossRef]
43. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2018**, *28*, 91–99. [CrossRef] [PubMed]
44. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *arXiv* **2016**, arXiv:1605.06409.
45. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
46. Han, J.; Ding, J.; Xue, N.; Xia, G.S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
47. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2021; Volume 35.
48. Zhao, P.; Qu, Z.; Bu, Y.; Tan, W.; Guan, Q. Polardet: A fast, more precise detector for rotated target in aerial images. *Int. J. Remote Sens.* **2021**, *42*, 5831–5861. [CrossRef]
49. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
50. Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; Ye, Q. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8792–8801.
51. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1829–1838.

52. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3520–3529.
53. Qing, Y.; Liu, W.; Feng, L.; Gao, W. Improved Yolo network for free-angle remote sensing target detection. *Remote Sens.* **2021**, *13*, 2171. [CrossRef]
54. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749. [CrossRef]
55. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.S.; Bai, X. Rotation-sensitive regression for oriented scene text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
56. Wang, Y.; Zhang, Y.; Zhang, Y.; Zhao, L.; Sun, X.; Guo, Z. SARD: Towards scale-aware rotated object detection in aerial imagery. *IEEE Access* **2019**, *7*, 173855–173865. [CrossRef]
57. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 2849–2858.
58. Sun, P.; Zheng, Y.; Zhou, Z.; Xu, W.; Ren, Q. R4 Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image Vis. Comput.* **2020**, *103*, 104036. [CrossRef]

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.