remote sensing

MDPI

Article

YOLO-HR: Improved YOLOv5 for Object Detection in High-Resolution Optical Remote Sensing Images

Dahang Wan [1,2], Rongsheng Lu [1,2,*], Sailei Wang [1,2], Siyuan Shen [1,2], Ting Xu [1,2] and Xianli Lang [1,2]

[1] School of Instrument Science and Opto-Electronic Engineering, Hefei University of Technology, Hefei 230009, China
[2] Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, Hefei University of Technology, Hefei 230009, China
* Correspondence: rslu@hfut.edu.cn

Abstract: Object detection is essential to the interpretation of optical remote sensing images and can serve as a foundation for research into additional visual tasks that utilize remote sensing. However, the object detection network currently employed in optical remote sensing images underutilizes the output of the feature pyramid, so there remains potential for an improved detection. At present, a suitable balance between the detection efficiency and detection effect is difficult to attain. This paper proposes an enhanced YOLOv5 algorithm for object detection in high-resolution optical remote sensing images, utilizing multiple layers of the feature pyramid, a multi-detection-head strategy, and a hybrid attention module to improve the effect of object-detection networks for use with optical remote sensing images. According to the SIMD dataset, the mAP of the proposed method was 2.2% better than YOLOv5 and 8.48% better than YOLOX, achieving an improved balance between the detection effect and speed.

Keywords: object detection; remote sensing image; attention mechanism; large resolution image; feature reuse; deep learning

Citation: Wan, D.; Lu, R.; Wang, S.; Shen, S.; Xu, T.; Lang, X. YOLO-HR: Improved YOLOv5 for Object Detection in High-Resolution Optical Remote Sensing Images. *Remote Sens.* **2023**, *15*, 614. https://doi.org/10.3390/rs15030614

Academic Editor: Hossein M. Rizeei

Received: 1 December 2022
Revised: 8 January 2023
Accepted: 9 January 2023
Published: 20 January 2023

1. Introduction

With the rapid development of remote sensing technology, high-resolution optical remote sensing images have been utilized to depict numerous items on the Earth's surface, including aircrafts, automobiles, buildings, etc. [1]. Object detection plays a crucial role in the interpretation of remote sensing images and can be used for their segmentation [2,3], description [4,5], and target tracking [6]. However, aerial optical remote sensing images manifest a diversity of scale, viewpoint specificity, random orientation, and high background complexity due to their relatively large field of view and the necessity for a high altitude [7,8], whereas the majority of conventional datasets contain ground-level views. Therefore, the object-detection techniques used in the construction of artificial features traditionally have a poor track record for accuracy and speed. The target-detection algorithm based on a convolutional neural network is significantly more efficient and effective than traditional target-detection algorithms. Due to the needs of society and supported by the development of deep learning, the use of neural networks for target detection in optical remote sensing images is a necessity.

Current object-detection algorithms incorporating deep learning to analyze optical remote sensing photographs can be classified as supervised, poorly supervised, or unsupervised. However, due to the complexity and instability of unsupervised and weakly supervised algorithms, supervised algorithms are the most frequently used. Moreover, supervised object-detection algorithms can be classified as either single-stage or two-stage. For instance, the two-stage capsule network SAHR-CapsNet [9] can accurately detect targets in remote sensing pictures. Due to the comparatively late discovery of capsule networks, however, the vast majority of modern two-stage object-detection algorithms

have been based on the RCNN [10–12] series. The methods described by previous researchers [13] have integrated the techniques of dilation convolution [14] and OHEM [15] into Faster RCNN [12] frameworks to improve the detection precision of small objects with a high density in optical remote sensing images. A similar practice is described in reference [16]. Wang et al. [17] proposed a fine detector with contextual information to improve the region suggestion network (RPN) in Faster RCNN to overcome background clutter and difficulties in recognition of foreground items in remote sensing images. The detection of airports and ports in downsampled satellite images, followed by mapping the discovered items back to the original ultra-high resolution satellite images, can successfully enable the concurrent detection of objects of varying sizes, according to research [18], based on the assumption that airplanes are often located in airports and ships are in ports and oceans. Weng et al. [19] proposed a rotating object-detection approach based on RCNN [10] to improve the accuracy of object detection in remote sensing images by addressing the randomization of target orientation.

Although two-stage object-detection algorithms are comparatively advantageous in terms of the accuracy and detection effect, they tend to involve complex models and operate at a slower speed, so researchers have focused on single-stage algorithms allowing for a trade-off of the speed and accuracy, such as YOLO [20–25], SSD [26,27], MobileNet [28–30], etc., adding an attention mechanism that improves the network's ability to detect remote sensing images. Previous research [31] integrated the attention mechanism CBAM [32] into the lightweight YOLOX network [33] to improve its detection accuracy for small targets in remote sensing image datasets. Another study [34] replaced the backbone of YOLOv3 with the backbone of MobileNetV3. It combined the attention mechanism to develop a lightweight single-stage object detection network SeMo-YOLO with an increased detection speed in the remote sensing object-detection network. However, instead of using the output of FPN and PANET simultaneously, the aforementioned YOLO-based network uses only one. However, the feature pyramid's output has been underutilized, and the detection impact could still be enhanced. The single-stage object detection network is a dense anchor box network with the problem of a positive and negative sample imbalance [35], so previous researchers [36] added focal loss [35] into the training process to alleviate the positive and negative sample imbalance and used the self-attention mechanism to extract high-level semantic information from the depth-feature map for target spatial localization, thereby improving the accuracy of the SSD model [26] for target localization. The aforementioned single-stage target-detection technique, which is used in remote sensing photographs, detects targets more quickly than a two-stage target detection network as it fixes the network input to a particular size, such as 640 × 640 or 512 × 512. However, small targets (dozens of pixels or even fewer) tend to be lost when this method is employed for high-resolution remote sensing target-detection datasets, leading to the poor detection of small targets and a deterioration in the model's overall detection effect caused by a reduction in the resolution [36–38].

The aforementioned techniques are crucial for object detection in remote sensing images, yet the following issues remain.

(1) Some researchers employed a two-stage model for object recognition, which is characterized by a high model complexity, a large number of parameter calculations, and a sluggish performance.

(2) Some researchers used single-stage networks for object detection in optical remote sensing images. However, most of them are scaled down to a lower input resolution, which diminishes the effectiveness of model detection.

(3) The majority of YOLO-based networks utilize either the output of FPN [39] or PAFPN [40], but not both.

(4) Some researchers have demonstrated that the hybrid attention mechanism can enhance the precision of object detection in optical remote sensing images. Nevertheless, hybrid attention modules are scarce now.

We proposed a lightweight object detection network for high-resolution remote sensing images based on the YOLOv5 framework in order to balance the detection accuracy, speed, the number of model parameters, and the existing features. The following is a summary of the contributions of this paper:

(1) Based on the YOLOv5 network topology, a single-stage object detection network named YOLO-HR for high-resolution optical remote sensing photographs was suggested.

(2) A multi-detection-head approach that can exploit the features of both FPN and PANET was proposed.

(3) A lightweight hybrid attention module was proposed.

(4) The model's efficiency and viability were validated using the SIMD [41] dataset.

The article is structured as follows: Section 2 presents a brief summary of the related works from the remote sensing object detection dataset, object detection network, and attention mechanism. Based on an analysis of the existing detection head output strategy in single-stage algorithms, Section 3 offers a multi-detection-head strategy and the YOLO-HR network. Section 4 evaluates the proposed technique on the SIMD datasets and presents the experimental results. The paper is summarized in Section 5.

2. Related Work

2.1. Datasets of Optical Remote Sensing Image Object Detection

Traditional remote sensing image datasets for object detection are reviewed in Table 1. Where Dataset represents the name of the dataset, Categories denotes the number of categories in the dataset, Images is the number of images, Instances represents the total number of targets, and Year represents the release year of the dataset. The recently published SIMD [41] dataset was utilized for this study, with the majority of its images measuring 1024 by 768 pixels.

Table 1. Universal remote sensing image object detection dataset.

DataSet	Categories	Images	Instances	Year
TAS [42]	1	30	1319	2008
SZTAKI-INRIA [43]	1	9	665	2012
NWPU VHR-10 [44]	10	800	3775	2014
VEDAI [45]	9	1210	3640	2015
UCAS-AOD [46]	2	910	6029	2015
DLR-MVDA [47]	2	20	14,235	2015
HRSC-2016 [48]	1	1070	2976	2016
RSOD [1]	4	976	6950	2017
DOTA [49]	15	2806	188,282	2017
DIOR [1]	20	23,463	192,472	2018
LEVIR [50]	3	21,952	10,069	2018
ITCVD [51]	1	173	29,088	2018
SIMD [52]	15	5000	45,303	2020

2.2. Attention Mechnishem

Currently, attention mechanisms in deep learning are commonly categorized as soft attention, hard attention, and self-attention. The soft attention mechanism assigns a weight between 0 and 1 to each input item and evaluates the majority of the data, but not equally. The hard attention mechanism assigns a weight of 0 or 1 to each input item. Unlike soft attention, hard attention just considers the component that demands attention and promptly discards unnecessary information. The self-attention mechanism assigns a weight to each input item based on the interaction between the input items, i.e., the "voting" between the input items determines which input items receive attention. The soft attention method is the most widespread in the field of remote sensing image object detection, and its representative articles include SE [53], CBAM [32], ECA [54], Co-Attention [55], Reverse Attention [56], Cross Attention [57], etc. Numerous articles [31,58–62] have demonstrated that mixed attention mechanisms improved the effect of a remote sensing target detection network,

including strengthening the detection effect and increasing the detection accuracy. In this study, the hybrid attention module MAB also consisted of hybrid soft attention mechanisms.

2.3. Object Detection Networks in Remote Sensing Image

Radar-based object detection and optical remote sensing object detection are the two types of remote sensing picture object detection. Optical sensors require favorable weather conditions and ample sunshine to produce high-quality photographs. The most notable advantage of radar sensors is that they are unaffected by the weather [63]. For example, the synthetic aperture radar (SAR) is a high-resolution image radar that can detect camouflage and penetrate masking objects in all-weather situations. One of the current hot topics in remote sensing is the application of neural networks to detect SAR images with complex and variable scenes [63–66]. This paper focuses on object detection in optical remote sensing images. Typically, image data are often derived from satellite photographs, such as Google Earth, or aerial images, such as UAS. The recent applications of deep learning to recognize objects in optical remote sensing images have produced satisfactory results. Wang et al. [67–69] took advantage of the advancements in the Faster RCNN [12], RetinaNet [35], and YOLOv3 [22] networks to detect wildlife in high-resolution UAV images. Sun et al. proposed a comprehensive partial-based convolutional neural network called PBNet for composite object detection in high-resolution optical remote sensing images [70]. In the past, RCNN was used to identify aircraft targets in very high-resolution remote sensing photographs with a poor precision and sluggish speed. Therefore, a mix of dense convolutional networks, multi-scale representation methods, and a number of enhancement techniques were utilized to strengthen the fundamental VGG16-Net's structure, raise accuracy, and more effectively recognize the target in satellite optical remote sensing images. [13]. The experiments mentioned [13,35,63–70] above used horizontal boundary boxes (HBB), which sometimes do not offer precise direction and scale information and have an excessive number of superfluous pixels in the backdrop. In addition, HBB and non-maximal inhibition (NMS) collaboration usually leads to missing detection when detecting objects with high aspect ratios and dense parking. In recent years, the recognition of directional objects (OBB) in RS images has garnered growing attention [71–76]. OBBs are often slower than HBBs in training and deployment. Hence, HBBs remain the focus of the current research. The algorithm of this study was also based on HBB.

The majority of those above high-resolution optical remote sensing target detection algorithms, which are typically classified into single-stage and double-stage target detection networks, are based on the existing mainstream target detection networks. RCNN [10], SPPNet [77], Fast RCNN [11], Faster RCNN [12], Cascade RCNN [78], etc., are examples of two-stage ones. The single-stage products include RetinaNet [35], SSD [26,27], FCOS [79], CenterNet [80], CornerNet [81], YOLOv1–7 [20–25,82], YOLOx [33], and YOLOF [83]. Traditional lightweight online models such as MobileNet [28–30], ShuffleNet [84,85], and Efficientdet [86] are also available.

3. Materials and Methods

3.1. YOLO-HR

3.1.1. Comparison of Prediction Head

The majority of the current YOLO series detection heads are based on the output feature of FPN and PAFPN, where FPN-based networks such as YOLOv3 and its variants are shown in Figure 1a, which directly utilize the one-way fused features for the output, and the PAFPN-based algorithms of YOLOv4 and YOLOv5 add a low-level to high-level channel on top of this, which directly transmits the low-level information upwards (Figure 1b). As demonstrated in Figure 1c and similarly in some studies [87–89], Zhu et al. added a detection head for a particular detection task in the TPH-YOLOv5 model. In Figure 1b,c, only the PAFPN features are used for the output, while the FPN features are underutilized. Therefore, YOLOv7 attaches three auxiliary heads to the FPN output, as depicted in

Figure 1d, although the auxiliary heads are only used for a "rough selection" and have a low weight. The detecting head of SSD was proposed to improve the too-coarse design of the anchor set by the YOLO network, as depicted in Figure 1e, and the design concept consists mostly of a dense anchor design with multiple aspect ratios at multiple scales. Inspired by Figure 1c–e, this paper proposed a multi-detection-head strategy for the YOLO detection head, as depicted in Figure 1f, which could utilize the feature information of PANet and FPN simultaneously. Additionally, an output head was added directly at the 64-fold downsampling, which caused the network to contain the prior global information.

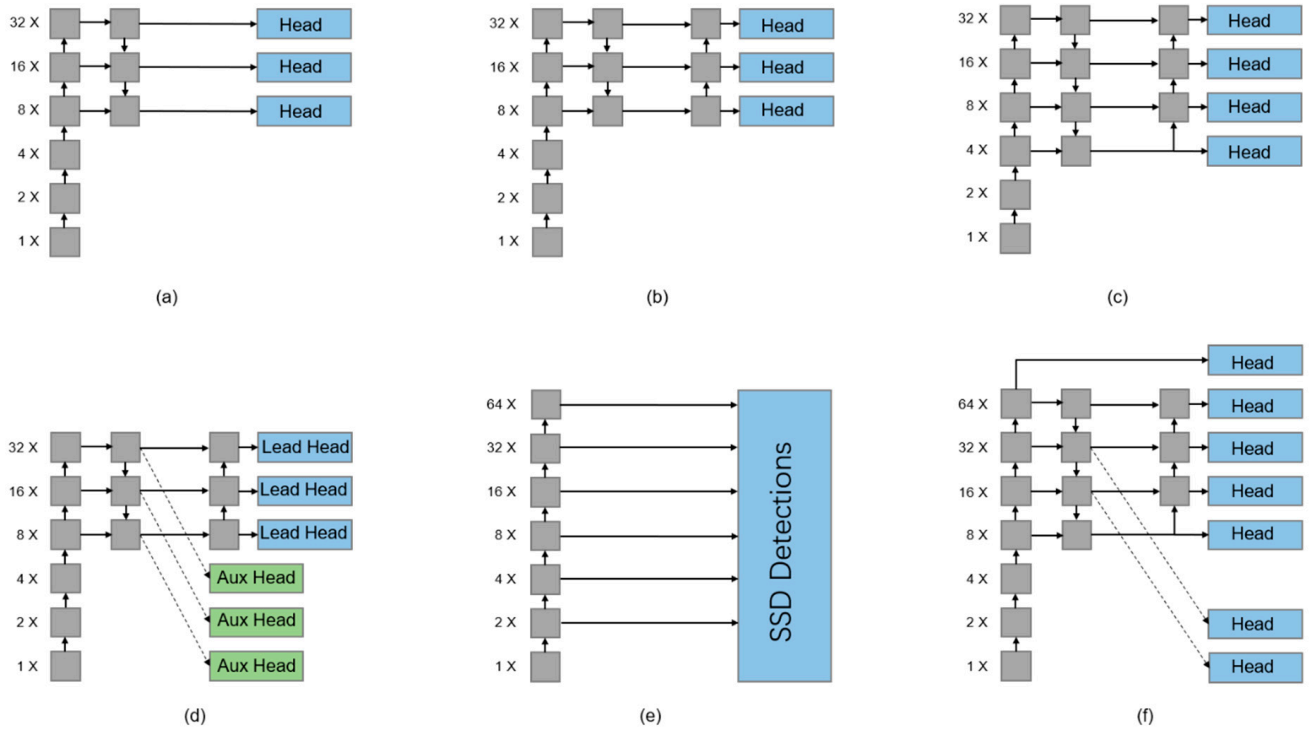


Figure 1. Comparison of the output of various detection heads. (**a**) FPN-based, (**b**) PANet-based, (**c**) TPH-YOLOv5, (**d**), Lead Head + Auxiliary Head, (**e**), SSD-based, (**f**) YOLO-HR (Ours).

3.1.2. Overall Structure of YOLO-HR

The multi-detection-head method could efficiently use the network's output features. YOLO-HR was an object detection network for high resolution remote sensing photographs. As depicted in Figure 2, the YOLO-HR network described in this paper can be separated into Backbone, Neck, and Head. The basic structure of Backbone was a CSP-DenseNet with C3 and Convolutional modules at its core. After the data enhancement, images were fed into the network and numerous convolutional modules retrieved features after channel mixing by the Conv module with a kernel size = 6. They were connected to PANet in Neck after the feature enhancement module named SPPF. Bidirectional feature fusion was undertaken to enhance the network's detecting capability. Conv2d was used to independently scale the fused feature layers to generate the multi-layer outputs. As depicted in Figure 3a, the NMS algorithm combined the outputs of all single-layer detectors to produce the final detection frame.

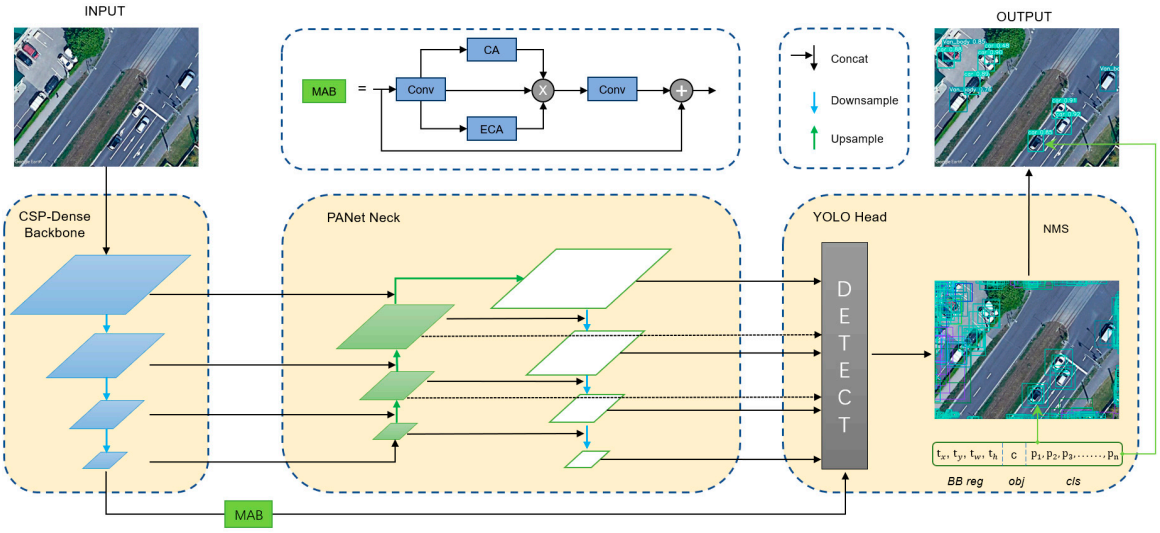


Figure 2. The overall structure of YOLO-HR.

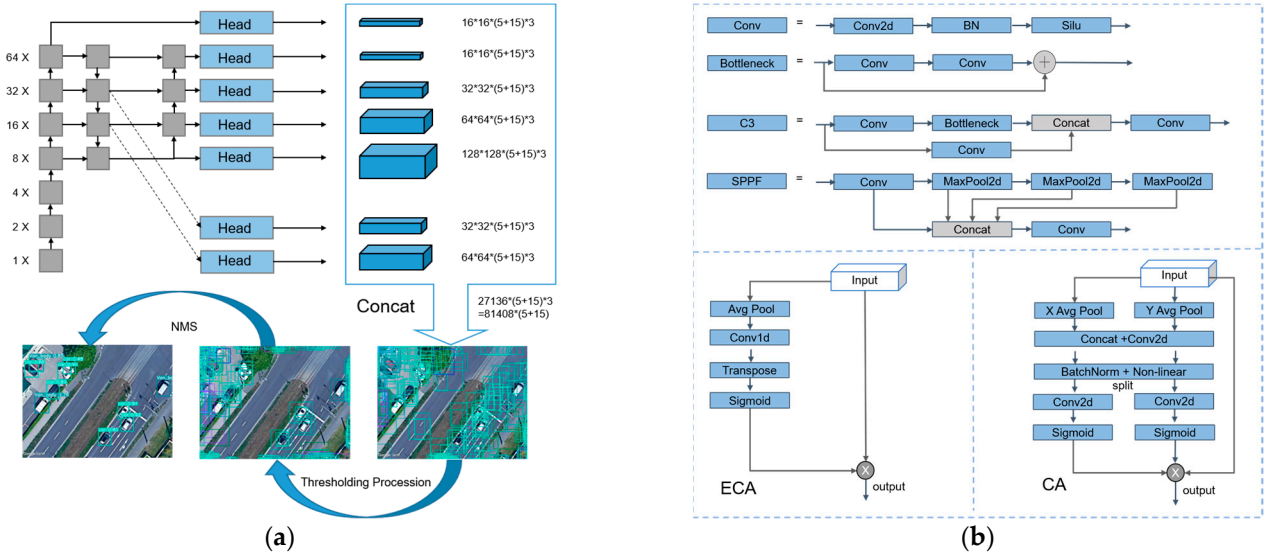


Figure 3. Composition modules of YOLO-HR. (**a**) The principle of YOLO-HR multi-head output; (**b**) the other composition modules of YOLO-HR.

Figure 3b depicts the structural composition of each module of the YOLO-HR network. Conv comprises a 2D convolutional layer, BN layer batch normalization, and Silu activation function, C3 comprises two 2D convolutional layers plus a bottleneck layer, and Upsample is the upsampling layer. The SPPF module is a sped-up version of the SPP module, and the MAB module is depicted in Figure 2, where the ECA [54] is depicted in the bottom left corner. After channel-level global average pooling without dimension reduction, the ECA is efficiently performed using the rapid 1D convolution of size k to capture local cross-channel interaction information, taking into account each channel's relationship with its k neighbors. The CA attention mechanism [55] is depicted in Figure 1's lower right corner, which encodes each channel along the horizontal and vertical coordinates, respectively, using a channel-level global average pooling of size (H,1) or (1, W) pooling kernel. The above two transformations collect features along two spatial directions to produce a pair of direction-aware feature maps, which are then concatenated and modified with convolution and Sigmoid functions to provide the attention output.

Table 2 displays the parameter settings for the entire network's structure. Input displays the input size of the image, Output displays the output size of the current layer,

Argvs are the input parameters of the current module, From represents the input source of the current layer, N represents the number of repetitions of the current module, and Parameters displays the size of the parameter number of the current layer.

Table 2. Parameter setting of the network structure.

ID	Module	From	N	Argvs	Output	Parameters
Input					1024 × 1024 × 3	
0	Conv	−1	1	(3, 32, 6, 2, 2)	512 × 512 × 32	3520
1	Conv	−1	1	(32, 64, 3, 2)	256 × 256 × 64	18,560
2	C3	−1	1	(64, 64, 1)	256 × 256 × 64	18,816
3	Conv	−1	1	(64, 128, 3, 2)	128 × 128 × 128	73,984
4	C3	−1	2	(128, 128, 2)	128 × 128 × 128	115,712
5	Conv	−1	1	(128, 256, 3, 2)	64 × 64 × 256	295,454
6	C3	−1	3	(256, 256, 3)	64 × 64 × 256	625,152
7	Conv	−1	1	(256, 384, 3, 2)	32 × 32 × 384	885,504
8	C3	−1	1	(384, 384, 1)	32 × 32 × 384	665,856
9	Conv	−1	1	(384, 512, 3, 2)	16 × 16 × 512	1,770,496
10	C3	−1	1	(512, 512, 1)	16 × 16 × 512	1,182,720
11	SPPF	−1	1	(512, 512, 5)	16 × 16 × 512	656,896
12	Conv	−1	1	(512, 384, 1, 1)	16 × 16 × 374	197,376
13	Upsample	−1	1	(None, 2, ‘nearest’)	32 × 32 × 384	0
14	Concat	(−1, 8)	1	(1)	32 × 32 × 768	0
15	C3	−1	1	(768, 384, 1, False)	32 × 32 × 384	813,312
16	Conv	−1	1	(384, 256, 1, 1)	32 × 32 × 256	98,816
17	Upsample	−1	1	(None, 2, ‘nearest’)	64 × 64 × 256	0
18	Concat	(−1, 6)	1	(1]	64 × 64 × 512	0
19	C3	−1	1	(512, 256, 1, False)	64 × 64 × 256	361,984
20	Conv	−1	1	(256, 128, 1, 1)	64 × 64 × 128	33,024
21	Upsample	−1	1	(None, 2, ‘nearest’)	128 × 128 × 128	0
22	Concat	(−1, 4)	1	(1)	128 × 128 × 256	0
23	C3	−1	1	(256, 128, 1, False)	128 × 128 × 128	90,880
24	Conv	−1	1	(128, 128, 3, 2)	64 × 64 × 128	147,712
25	Concat	(−1, 20)	1	(1)	64 × 64 × 256	0
26	C3	−1	1	(256, 256, 1, False)	64 × 64 × 256	296,448
27	Conv	−1	1	(256, 256, 3, 2)	32 × 32 × 256	590,336
28	Concat	(−1, 16)	1	(1)	32 × 32 × 512	0
29	C3	−1	1	(512, 384, 1, False)	32 × 32 × 384	715,008
30	Conv	−1	1	(384, 384, 3, 2)	16 × 16 × 384	1,327,872
31	Concat	(−1, 11)	1	(1)	16 × 16 × 896	0
32	C3	−1	1	(896, 512, 1, False)	16 × 16 × 512	1,379,328
33	MAB	11	1	(512, 512)	16 × 16 × 512	1,361,477
34	Detect			(23, 26, 20, 16, 29, 32, 33)	81406 × (5 + 15)	

3.2. Data Augmentation

The essence of data augmentation is to artificially introduce human visual prior knowledge, which can improve the performance of the model very well, and it has basically become the standard for model training. The more commonly used geometric transformation methods are flip, rotate, crop, scale, pan, dither, etc. The pixel transformation methods include adding pretzel and Gaussian noise, performing a Gaussian blur, adjusting the HSV contrast, and adjusting the brightness, saturation, histogram equalization, white balance, etc. In addition to the above methods, this paper also uses a variety of data enhancement methods in the training phase, each with different random ratios, such as Mosica, CUTOUT, small target replication, etc. Among them, flip and rotation are used to solve the problem of Angle diversity in remote sensing images, zoom and shift are used to solve the problem of the multi-scale in remote sensing images, dithering and adding noise are used to improve the problem of a complex background in remote sensing images, and small target replication is used to expand the samples and improve the detection effect of small targets.

3.3. Loss Function

The loss function of YOLO-HR was composed of three components: target confidence loss, target category loss, and target positioning loss. The loss function could be expressed as follows:

$$L_{all} = \lambda_{conf} L_{conf} + \lambda_{cls} L_{cls} + \lambda_{loc} L_{loc} \tag{1}$$

where L_{all} contained three hyperparameters, the weight of each component, which could be modified before training based on the actual circumstances. In this work, the corresponding weights of the three sections were 1.0, 0.5, and 0.05. The target confidence loss utilized the BCE (binary cross-entropy) loss, with the following expression:

$$\begin{aligned} L_{conf} = &-\sum_{i=1}^{K*K}\sum_{j=1}^{B} {I_{ij}}^{obj}\left[C_i^j log(C_i'^j) + \left(1 - C_i^j\right) log(1 - C_i'^j)\right] \\ &-\lambda_{noobj}\sum_{i=1}^{K*K}\sum_{j=1}^{B} {I_{ij}}^{noobj}\left[C_i^j log(C_i'^j) + \left(1 - C_i^j\right) log(1 - C_i'^j)\right] \end{aligned} \tag{2}$$

Among them, $K * K$ could take on three distinct values, with the particular size being dependent on the image size. Taking 1024×1024 as an example, they were 16×16, 32×32, 64×64, and 128×128, respectively, illustrating the number of grids on the feature maps generated by YOLO-HR at three different scales. B represented the number of preceding boxes. ${I_{ij}}^{obj}$ specifies whether the jth previous box of the ith grid had a prediction target. ${I_{ij}}^{obj}$ is 1 if the condition was met; else, it was 0. ${I_{ij}}^{noobj}$ indicated if the jth previous box of the ith grid did not contain a predicted target. If not, ${I_{ij}}^{noobj}$ was 1; otherwise, it was 0. ${C_i}^j$ and $C_i'^j$ represented the actual and expected confidence values, respectively. λ_{noobj} was a constant coefficient, typically assumed to be 0.5, that was used to balance the positive and negative samples.

The target confidence loss was also the BCE loss, and the expression is as follows:

$$L_{cls} = -\sum_{I=0}^{K*K} {I_{ij}}^{obj} \sum_{c \in classes} \left\{ P_i^j(c) log[P_i'^j(c)] + \left[1 - P_i^j(c)\right] log[1 - P_i'^j(c)] \right\} \tag{3}$$

where $K * K$, B and ${I_{ij}}^{obj}$ were consistent with Equation (1), c was the target category, and $P_i^j(c)$ and $P_i'^j(c)$ were the probability that the target in the jth prediction box in the ith grid belongs to the real value and the predicted value of a certain category, respectively.

The SIoU loss [90] replaced the CIoU loss [91] function for the target positioning loss in order to increase the training speed and reasoning precision in this paper. The following is the formula:

$$L_{box} = 1 - IoU + \frac{\Omega + \Delta}{2} \tag{4}$$

$$IoU = \frac{\left|B \cap B^{GT}\right|}{\left|B \cup B^{GT}\right|} \tag{5}$$

$$\Delta = \sum_{t=x,y} 1 - e^{-\gamma \rho_t} \tag{6}$$

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega_t}\right)^{\theta} \tag{7}$$

where Δ represented the Distance cost, γ the Angle cost, Ω the Shape cost, and θ expressed the Shape cost level of concern.

4. Experiment

4.1. Experimental Platform and Related Indexes

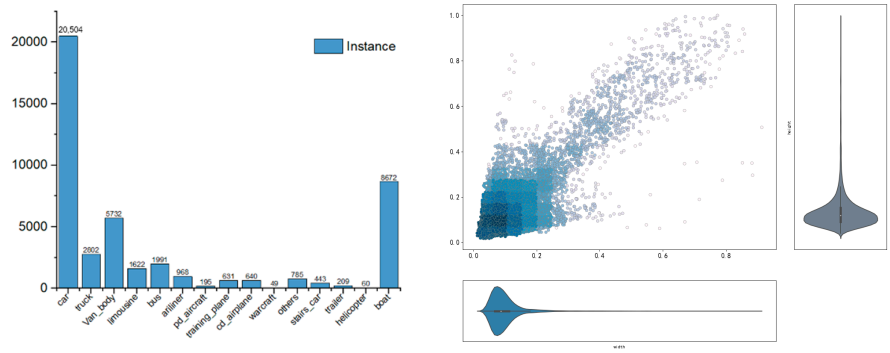
4.1.1. Experimental Platform

The experimental platform of this paper is shown in Table 3.

Table 3. Experimental platform.

Platform	Name
CPU	lntel(R) Core(TM) i9-12900K/32G
GPU	NVIDIA GeForce RTX 3090/24G
Disk capacity	SSD/500G + HDD/4T
The operating system	Windows 10
Deep learning framework	Pytorch 1.7

The SIMD dataset is a multi-category, open-source, high-resolution remote sensing object detection dataset containing a total of 15 classes, as illustrated in Figure 4. Additionally, the SIMD dataset is more distributed with small- and medium-sized targets (w < 0.4, h < 0.4), and the detection head used by YOLO-HR proposed in this paper to detect this region is double the number of detection heads used by the common YOLO algorithm, so YOLO-HR has greater advantages on this dataset.



(**a**) (**b**)

Figure 4. Distribution of targets in SIMD dataset. (**a**) shows the distribution of the number of categories; (**b**) shows the distribution of target width and height in the image; the color from white to blue (from light to dark) indicates a more concentrated distribution.

4.1.2. Related Indexes

The network performance evaluation is mainly based on the mAP (mean average accuracy) during training and the performance of the trained network in the validation set. To measure the detection results quantitatively, the accuracy Precision, Recall, and mAP are used here as the performance evaluation metrics. The expressions of P and R are as follows.

$$\begin{cases} Recall = TP/(TP + FN) \\ Precision = TP/(TP + FP) \end{cases} \tag{8}$$

where True positives (TP) are the number of samples that are actually positive and classified as positive by the classifier; True negatives (TN) are the number of samples that are actually negative and classified as negative by the classifier; False positives (FP) are the number of samples that are actually negative but classified as positive by the classifier; and False negatives (FN) are the number of samples that are actually positive but classified as negative by the classifier.

Average Precision (AP) is the area enclosed by the P-R curve. Usually, the better the classifier, the higher the AP value. Mean Average Precision (mAP) is the AP of each category taken separately, and then the average of the AP of all categories is calculated, representing a composite measure of the average precision of the detected targets. AP50 in the later text means that the IoU threshold is greater than 0.5, mAP is mAP 0.5:0.95, and step is 0.05.

4.2. Experiments on SIMD

4.2.1. Ablation Test

It was possible to connect the output of the SPPF module to the output head and thus identify large targets in the image. However, the output of the SPPF module had multiple connections and is concerned with targets of multiple scales, so using it directly for the detection head to identify large objects would result in a poor model representation, as shown in Figure 5. Figure 5 depicts a visual comparison of the heat map of some detection findings prior to and following the addition of the MAB module. After adding the MAB module, this detection head focused on detecting large objects, while the prediction of small targets was assigned to other prediction heads and the expression effect of the model was improved, which was also more in line to divide the detection head based on the target size in the YOLO algorithm.

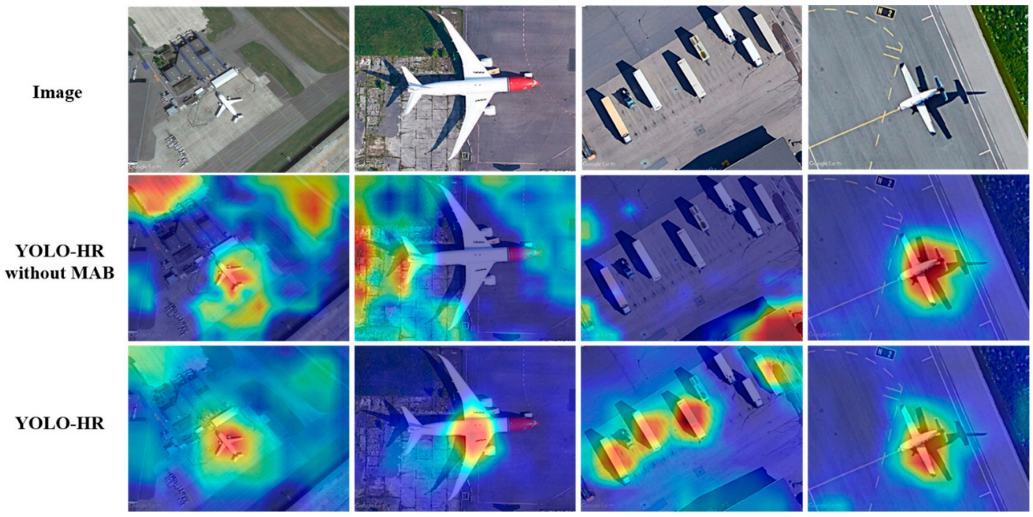


Figure 5. Heat map visualization of the partial detection results before and after adding the MAB module. Both are visualized with Grad-CAM [92].

Using the calculation results of the YOLOv5s algorithm as a reference, the effects of the MPH output strategy and MAB module on the calculation results were examined in the SIMD dataset and 1024 × 1024 image resolution, as shown in Table 4 and Figure 6, respectively, from top to bottom, indicating that the increase in the modules is in order. Finetune means the model was pre-trained on the ImageNet dataset, and then the trained model was fine-tuned on the SIMD dataset. The results showed that after the addition of the MPH strategy and MAB module, the number of parameters in the model increased by 2.5 M. Still, the increase in the number of parameters was negligible compared with the disk capacity of hundreds of G. The speed was not significantly improved, but the AP50 of the model increased by 2.1%, mAP increased by 2.2%, and the accuracy increased by 1.5%. The recall rate increased by 1.19%.

Table 4. Performance improvement of each part design on the result.

Name	Params (M)	FLOPS (G)	Speed (ms)	AP50 (%)	mAP (%)	P (%)	R (%)
YOLOv5s	6.72	15.9	5.8	81.51	62.8	75.01	79.66
YOLOv5s + Finetune	6.72	15.9	5.8	83.85	66.05	83.10	79.65
YOLOv5s +MPH	11.9	16.3	6.5	82.96	65.12	80.73	79.40
YOLOv5s + MPH + Finetune	11.9	16.3	6.5	85.15	66.68	83.47	80.30
YOLO-HR	13.2	16.6	6.7	83.61	65.0	76.51	80.85
YOLO-HR + Finetune	13.2	16.6	6.7	85.59	67.31	85.95	81.28

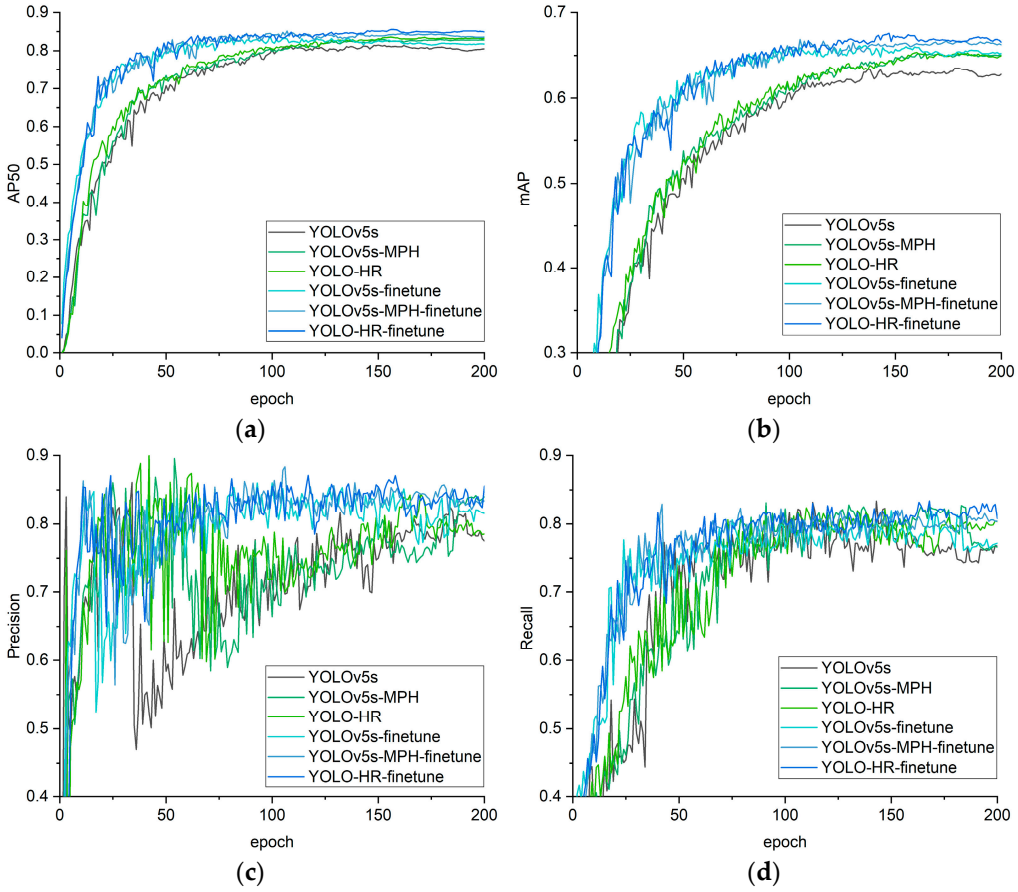


Figure 6. Results comparison chart. (**a**) AP50, (**b**) mAP, (**c**) Precision, (**d**), Recall.

4.2.2. Comparison Experiments

Simultaneously, the classic YOLOv3-Tiny, Faster RCNN, YOLOv7, and YOLOX models were selected for comparison tests in this paper. The Yolov3-Tiny, YOLOv5 (DenseNet + PAN), and YOLO-HR codes and pre-training models utilized in this experiment were obtained through the YOLOv5 open-source framework and the YOLOv7 models through the YOLOv7 open-source framework. The YOLOX(Darknet-53 + FPN) algorithm was derived from the literature [33], while the other models, including Faster RCNN(Resnet-50 + FPN), were derived from the MMDetection [93] open-source framework. We tested and compared YOLO-HR and other algorithms at a 1024 by 1024 image resolution. We merely compared the number of parameters to prevent variations caused by the model storage methods of various formats. For instance, the amount of parameters in YOLOv5s is 7.11 M, but Pytorch's model storage format is 14.4 M. In order to rule out randomness, the running time was computed as the average time for testing 1000 photos, as shown in Table 5. The suggested approach outperformed YOLOv5, YOLOv3-Tiny, YOLOv7-Tiny, YOLOX, and the Faster RCNN model using Resnet-50 as its backbone in terms of the detection outcomes (mAP and AP50). Although slightly more sophisticated than the YOLOv5 and YOLOX models, the number of references of a few meters was minimal, even compared to the modest storage space of edge devices such as Nvidia TX2 and NX, which was only 32 gigabytes, so it was more than sufficient. In terms of the speed, it was superior to YOLOv3-Tiny and Faster RCNN and the detection speed was only 0.5 ms higher than that of YOLOv5, without a substantial reduction in the detection speed. The complete detection findings of YOLO-HR proposed in this paper offered benefits over the appeal algorithm. The results of the experiments indicated that the YOLO-HR algorithm struck a more suitable balance between the reference number, speed, and detection effectiveness.

Table 5. Comparison with other algorithms.

Name	AP50 (%)	mAP (%)	Params (M)	FLOPs (G)	Speed (ms)
YOLOv3-Tiny [32,34]	77.23	54.53	8.3	12.9	4.3
YOLOv5n [24]	79.56	60.69	1.7	4.2	4.8
YOLOv5s [24]	83.85	66.05	6.72	15.9	5.8
YOLO-HR-n	83.01	64.04	3.34	4.4	6.3
YOLO-HR-s	85.59	67.31	13.2	16.6	6.7
YOLOX-s [33]	76.63	56.83	8.94	26.79	5.7
YOLOv7s [25]	83.80	66.55	8.92	26.8	6.3
YOLOv7-tiny [25]	82.08	64.16	5.77	13.1	5.2
Faster RCNN [12]	77.74	-	41.19	198.47	26.3

Some of the detection results are shown in Figure 7. From each detection result, there was not much difference with other algorithms, but compared with other algorithms, the algorithm in this paper improved the detection effect of the model while ensuring no significant increase in the time consumption and enhanced the expression effect of the model by using the attention mechanism.

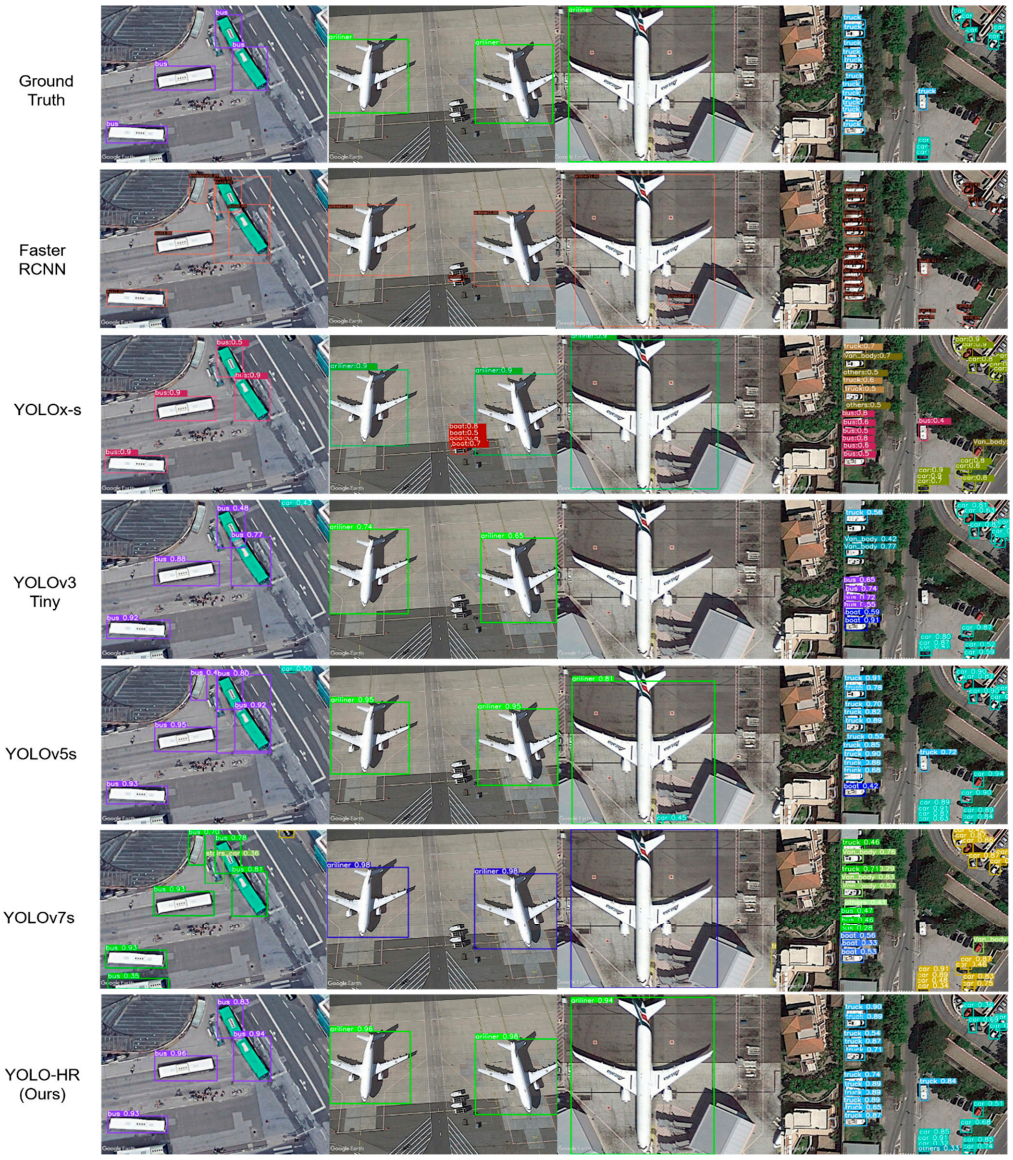


Figure 7. Comparison of the detection effect of different model detection parts.

4.2.3. Qualitative Results

Some qualitative results of the YOLO-HR algorithm proposed in this paper on the SIMD dataset are shown in Figure 8. As shown in the figure, the YOLO-HR model could better detect objects in remote sensing images with special viewing angles, including objects with complex backgrounds, random directions, and different scales.

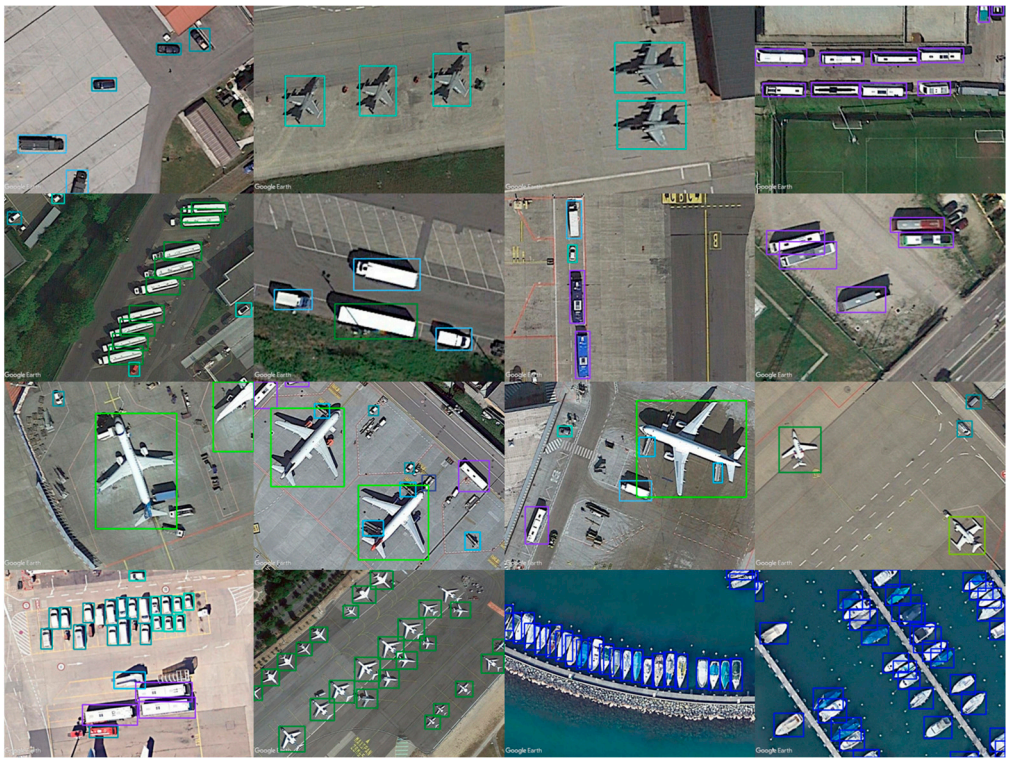


Figure 8. Some detection results on the SIMD dataset of YOLO-HR.

5. Conclusions

To address the issue that the majority of the current models utilized for optical remote sensing image object detection underutilized the output features of the feature pyramid, we proposed a multi-head strategy based on prior work and we proposed a hybrid attention module, MAB, for the lack of hybrid attention mechanisms. Finally, we embedded the aforementioned two methods into the YOLOv5 network and presented a high resolution optical remote sensing target recognition algorithm named YOLO-HR. The YOLO-HR algorithm employed several detection heads for object detection and recycled the output features of the feature pyramid, allowing the network to enhance the detection effect further. The experiments indicate that the YOLO-HR algorithm allows for a greater number of downsampling multiples and faster detection results than other algorithms while preserving the original detection speed. In subsequent work, we plan to extend and apply the concept of modifying the network structure presented in this paper to other object detection algorithms, study other feature reuse strategies, and investigate the deployment and application issues of the algorithm presented in this paper in greater depth.

Author Contributions: D.W. and R.L. conceived and designed the experiments; D.W. and S.W. performed the experiments; T.X. and S.S. analyzed the data; X.L. contributed analysis tools; D.W. wrote the paper; R.L. supervised this work. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 51875164); National Key Research and Development Program of China (No. 2018YFB2003801).

Data Availability Statement: The datasets presented in this study are available through: https://github.com/ihians/simd (accessed on 12 August 2022).

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

References

1. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark 2019. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
2. Bello, I.M.; Zhang, K.; Su, Y.; Wang, J.; Aslam, M.A. Densely Multiscale Framework for Segmentation of High Resolution Remote Sensing Imagery. *Comput. Geosci.* **2022**, *167*, 105196. [CrossRef]
3. 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]
4. Bonannella, C.; Chirici, G.; Travaglini, D.; Pecchi, M.; Vangi, E.; D'Amico, G.; Giannetti, F. Characterization of Wildfires and Harvesting Forest Disturbances and Recovery Using Landsat Time Series: A Case Study in Mediterranean Forests in Central Italy. *Fire* **2022**, *5*, 68. [CrossRef]
5. Li, J.; Zhuang, Y.; Dong, S.; Gao, P.; Dong, H.; Chen, H.; Chen, L.; Li, L. Hierarchical Disentangling Network for Building Extraction from Very High Resolution Optical Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 1767. [CrossRef]
6. Wu, D.; Song, H.; Fan, C. Object Tracking in Satellite Videos Based on Improved Kernel Correlation Filter Assisted by Road Information. *Remote Sens.* **2022**, *14*, 4215. [CrossRef]
7. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
8. Li, C.; Cong, R.; Guo, C.; Li, H.; Zhang, C.; Zheng, F.; Zhao, Y. A Parallel Down-up Fusion Network for Salient Object Detection in Optical Remote Sensing Images. *Neurocomputing* **2020**, *415*, 411–420. [CrossRef]
9. Yu, Y.; Wang, J.; Qiang, H.; Jiang, M.; Tang, E.; Yu, C.; Zhang, Y.; Li, J. Sparse Anchoring Guided High-Resolution Capsule Network for Geospatial Object Detection from Remote Sensing Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *104*, 102548. [CrossRef]
10. 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.
11. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
13. Ding, P.; Zhang, Y.; Deng, W.-J.; Jia, P.; Kuijper, A. A Light and Faster Regional Convolutional Neural Network for Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
14. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.
15. Training Region-Based Object Detectors with Online Hard Example Mining | IEEE Conference Publication | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/7780458 (accessed on 24 October 2022).
16. Shivappriya, S.N.; Priyadarsini, M.J.P.; Stateczny, A.; Puttamadappa, C.; Parameshachari, B.D. Cascade Object Detection and Remote Sensing Object Detection Method Based on Trainable Activation Function. *Remote Sens.* **2021**, *13*, 200. [CrossRef]
17. Wang, Y.; Xu, C.; Liu, C.; Li, Z. Context Information Refinement for Few-Shot Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3255. [CrossRef]
18. Wu, Z.-Z.; Wang, X.-F.; Zou, L.; Xu, L.-X.; Li, X.-L.; Weise, T. Hierarchical Object Detection for Very High-Resolution Satellite Images. *Appl. Soft Comput.* **2021**, *113*, 107885. [CrossRef]
19. Weng, L.; Gao, J.; Xia, M.; Lin, H. MSNet: Multifunctional Feature-Sharing Network for Land-Cover Segmentation. *Remote Sens.* **2022**, *14*, 5209. [CrossRef]
20. 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.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
22. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
23. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
24. Ultralytics/Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 15 September 2022).
25. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
26. 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 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9905, pp. 21–37. ISBN 978-3-319-46447-3.
27. Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.

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. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks 2019. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
30. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
31. Liu, K.; Huang, J.; Li, X. Eagle-Eye-Inspired Attention for Object Detection in Remote Sensing. *Remote Sens.* **2022**, *14*, 1743. [CrossRef]
32. 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.
33. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
34. Li, P.; Che, C. SeMo-YOLO: A Multiscale Object Detection Network in Satellite Remote Sensing Images. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; p. 8.
35. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection 2018. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
36. Wang, J.; Gong, Z.; Liu, X.; Guo, H.; Yu, D.; Ding, L. Object Detection Based on Adaptive Feature-Aware Method in Optical Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3616. [CrossRef]
37. Han, W.; Li, J.; Wang, S.; Wang, Y.; Yan, J.; Fan, R.; Zhang, X.; Wang, L. A Context-Scale-Aware Detector and a New Benchmark for Remote Sensing Small Weak Object Detection in Unmanned Aerial Vehicle Images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102966. [CrossRef]
38. Wang, Y.; Bashir, S.M.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]
39. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
40. 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.
41. Multisized Object Detection Using Spaceborne Optical Imagery | IEEE Journals & Magazine | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/9109702 (accessed on 10 November 2022).
42. Heitz, G.; Koller, D. Learning Spatial Context: Using Stuff to Find Things. In *Proceedings of the Computer Vision—ECCV 2008*; Forsyth, D., Torr, P., Zisserman, A., Eds.; Springer: Berlin, Heidelberg, 2008; pp. 30–43.
43. MPLab Earth Observation. Available online: http://web.eee.sztaki.hu/remotesensing/building_benchmark.html (accessed on 1 November 2022).
44. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images | IEEE Journals & Magazine | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/7560644 (accessed on 1 November 2022).
45. Razakarivony, S.; Jurie, F. Vehicle Detection in Aerial Imagery: A Small Target Detection Benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
46. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation Robust Object Detection in Aerial Images Using Deep Convolutional Neural Network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
47. DLR—Earth Observation Center—DLR Multi-Class Vehicle Detection and Orientation in Aerial Imagery (DLR-MVDA). Available online: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12760/22294_read-52777 (accessed on 1 November 2022).
48. Liu, Z.; Yuan, L.; Weng, L.; Yiping, Y. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017. [CrossRef]
49. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
50. Zou, Z.; Shi, Z. Random Access Memories: A New Paradigm for Target Detection in High Resolution Aerial Remote Sensing Images. *IEEE Trans. Image Process.* **2018**, *27*, 1100–1111. [CrossRef]
51. ITCVD Dataset—University of Twente Research Information. Available online: https://research.utwente.nl/en/datasets/itcvd-dataset (accessed on 1 November 2022).
52. Scottish Index of Multiple Deprivation 2020. Available online: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/ (accessed on 2 September 2022).
53. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
54. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* **2020**, arXiv:1910.03151.

55. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
56. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse Attention for Salient Object Detection. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11213, pp. 236–252. ISBN 978-3-030-01239-7.
57. Lin, H.; Cheng, X.; Wu, X.; Yang, F.; Shen, D.; Wang, Z.; Song, Q.; Yuan, W. CAT: Cross Attention in Vision Transformer 2021. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo, Taipei City, Taiwan, 11–15 July 2022.
58. Qingyun, F.; Zhaokui, W. Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery. *Pattern Recognit.* **2022**, *130*, 108786. [CrossRef]
59. Hu, G.; Yao, P.; Wan, M.; Bao, W.; Zeng, W. Detection and Classification of Diseased Pine Trees with Different Levels of Severity from UAV Remote Sensing Images. *Ecol. Inform.* **2022**, *72*, 101844. [CrossRef]
60. Song, C.; Zhang, F.; Li, J.; Xie, J.; Yang, C.; Zhou, H.; Zhang, J. Detection of Maize Tassels for UAV Remote Sensing Image with an Improved YOLOX Model. *J. Integr. Agric.* **2022**. *in press*. [CrossRef]
61. Wang, Z.; Wang, J.; Yang, K.; Wang, L.; Su, F.; Chen, X. Semantic Segmentation of High-Resolution Remote Sensing Images Based on a Class Feature Attention Mechanism Fused with Deeplabv3+. *Comput. Geosci.* **2022**, *158*, 104969. [CrossRef]
62. Lang, L.; Xu, K.; Zhang, Q.; Wang, D. Fast and Accurate Object Detection in Remote Sensing Images Based on Lightweight Deep Neural Network. *Sensors* **2021**, *21*, 5460. [CrossRef]
63. Zhao, C.; Fu, X.; Dong, J.; Qin, R.; Chang, J.; Lang, P. SAR Ship Detection Based on End-to-End Morphological Feature Pyramid Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4599–4611. [CrossRef]
64. Zhou, K.; Zhang, M.; Wang, H.; Tan, J. Ship Detection in SAR Images Based on Multi-Scale Feature Extraction and Adaptive Feature Fusion. *Remote Sens.* **2022**, *14*, 755. [CrossRef]
65. Han, P.; Liao, D.; Han, B.; Cheng, Z. SEAN: A Simple and Efficient Attention Network for Aircraft Detection in SAR Images. *Remote Sens.* **2022**, *14*, 4669. [CrossRef]
66. Yu, W.; Wang, Z.; Li, J.; Luo, Y.; Yu, Z. A Lightweight Network Based on One-Level Feature for Ship Detection in SAR Images. *Remote Sens.* **2022**, *14*, 3321. [CrossRef]
67. Peng, J.; Wang, D.; Liao, X.; Shao, Q.; Sun, Z.; Yue, H.; Ye, H. Wild Animal Survey Using UAS Imagery and Deep Learning: Modified Faster R-CNN for Kiang Detection in Tibetan Plateau. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 364–376. [CrossRef]
68. Torney, C.J.; Lloyd-Jones, D.J.; Chevallier, M.; Moyer, D.C.; Maliti, H.T.; Mwita, M.; Kohi, E.M.; Hopcraft, G.C. A Comparison of Deep Learning and Citizen Science Techniques for Counting Wildlife in Aerial Survey Images. *Methods Ecol. Evol.* **2019**, *10*, 779–787. [CrossRef]
69. Eikelboom, J.A.J.; Wind, J.; van de Ven, E.; Kenana, L.M.; Schroder, B.; de Knegt, H.J.; van Langevelde, F.; Prins, H.H.T. Improving the Precision and Accuracy of Animal Population Estimates with Aerial Image Object Detection. *Methods Ecol. Evol.* **2019**, *10*, 1875–1887. [CrossRef]
70. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-Based Convolutional Neural Network for Complex Composite Object Detection in Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 50–65. [CrossRef]
71. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240.
72. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; 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]
73. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented Objects as Pairs of Middle Lines. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 268–279. [CrossRef]
74. Wang, J.; Yang, W.; Li, H.-C.; Zhang, H.; Xia, G.-S. Learning Center Probability Map for Detecting Objects in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4307–4323. [CrossRef]
75. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
76. Zheng, X.; Zhang, W.; Huan, L.; Gong, J.; Zhang, H. AProNet: Detecting Objects with Precise Orientation from Aerial Images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 99–112. [CrossRef]
77. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
78. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
79. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
80. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
81. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
82. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.

83. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-Level Feature. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
84. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
85. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018. [CrossRef]
86. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
87. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 10–17 October 2021.
88. Kumar, A.; Kalia, A.; Kalia, A. ETL-YOLO v4: A Face Mask Detection Algorithm in Era of COVID-19 Pandemic. *Optik* **2022**, *259*, 169051. [CrossRef]
89. Li, J.; Gu, J.; Huang, Z.; Wen, J. Application Research of Improved YOLO V3 Algorithm in PCB Electronic Component Detection. *Appl. Sci.* **2019**, *9*, 3750. [CrossRef]
90. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
91. Chen, D.; Miao, D. Control Distance IoU and Control Distance IoU Loss Function for Better Bounding Box Regression. *arXiv* **2021**, arXiv:2103.11696.
92. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
93. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.