remote sensing

MDPI

Article

Improved One-Stage Detectors with Neck Attention Block for Object Detection in Remote Sensing

Kaiqi Lang [1,2], Mingyu Yang [1], Hao Wang [1], Hanyu Wang [1,2], Zilong Wang [1,2], Jingzhong Zhang [3] and Honghai Shen [1,*]

[1] Key Laboratory of Airborne Optical Imaging and Measurement, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Forest Protection Research Institute of Heilongjiang Province, Harbin 150040, China
* Correspondence: shenhh@ciomp.ac.cn

Abstract: Object detection in remote sensing is becoming a conspicuous challenge with the rapidly increasing quantity and quality of remote sensing images. Although the application of Deep Learning has obtained remarkable performance in Computer Vision, detecting multi-scale targets in remote sensing images is still an unsolved problem, especially for small instances which possess limited features and intricate backgrounds. In this work, we managed to cope with this problem by designing a neck attention block (NAB), a simple and flexible module which combines the convolutional bottleneck structure and the attention mechanism, different from traditional attention mechanisms that focus on designing complicated attention branches. In addition, Vehicle in High-Resolution Aerial Imagery (VHRAI), a diverse, dense, and challenging dataset, was proposed for studying small object detection. To validate the effectiveness and generalization of NAB, we conducted experiments on a variety of datasets with the improved YOLOv3, YOLOv4-Tiny, and SSD. On VHRAI, the improved YOLOv3 and YOLOv4-Tiny surpassed the original models by 1.98% and 1.89% mAP, respectively. Similarly, they exceeded the original models by 1.12% and 3.72% mAP on TGRS-HRRSD, a large multi-scale dataset. Including SSD, these three models also showed excellent generalizability on PASCAL VOC.

Keywords: remote sensing; multi-scale object detection; small object detection; attention mechanism; YOLOv3; YOLOv4-Tiny; SSD

Citation: Lang, K.; Yang, M.; Wang, H.; Wang, H.; Wang, Z.; Zhang, J.; Shen, H. Improved One-Stage Detectors with Neck Attention Block for Object Detection in Remote Sensing. *Remote Sens.* **2022**, *14*, 5805. https://doi.org/10.3390/rs14225805

Academic Editors: Yue Wu, Kai Qin, Qiguang Miao and Maoguo Gong

Received: 17 October 2022
Accepted: 14 November 2022
Published: 17 November 2022

1. Introduction

In remote sensing, multiple satellites and aircraft are used to capture images that contain significant information, such as the characteristics and changes of landscape, man-made targets, and traces. Object detection is a critical approach to extracting useful information from remote sensing images. It plays a vital role in environmental monitoring, geological hazard detection, land-use/land-cover mapping, geographic information system update, military reconnaissance and location, and land planning [1].

Traditional object detectors, which are usually composed of region proposal, feature extraction, feature fusion, and classifier training, require elaborately hand-made features and must be trained step by step. Therefore, these methods have inferior efficiency, accuracy, and generalizability. Especially with the rapid advancement of the quantity and quality of optical remote sensing images, these methods can not meet the requirement of practical applications by degrees.

In the last decades, convolutional neural networks (CNNs) have made tremendous breakthroughs in various computer vision tasks, including image classification, object detection, and semantic segmentation. The application of CNNs in object detection for remote sensing images achieves better accuracy, higher efficiency, and more powerful

generalizability than traditional methods. A common CNN detector is composed of a backbone, which is pretrained with large datasets and used to extract feature maps; a neck, which can enhance feature representation and make feature transition smooth from feature maps to output; and a head, which is used to generate regression and classification predictions.

The backbone, the most significant part of a CNN model, determines the fundamental performance of a CNN model. Since the advent of AlexNet [2], a variety of backbones have been designed for improving the capability of feature extraction, such as VGG16 [3], Inception [4], ResNet [5], ResNeXt [6], and Darknet53 [7]. In these backbones, an important research direction is to increase the depth and width of the network. AlexNet only has five convolutional layers, and VGG16 has sixteen convolutional layers. After the creation of a residual block, ResNet-152 contains 152 convolutional layers. Meanwhile, the set of Inception structures. which concentrates on increasing the width of a model, also obtains excellent performance.

The purpose of the neck is to refine feature maps from the backbone and transmit them to the head. In order to aggregate bottom and top features, a Feature Pyramid Network (FPN) [8] is designed to combine low-resolution features and high-resolution features by adding a top-down path. To address the shortcoming that top feature maps lack location information in FPN, a Path Aggregation Network (PAN) [9] further adds a down-top path on the basis of the FPN. Although the neck has a significant function in enhancing feature representation and making feature transition from feature maps to output smooth, the research for the neck is still inadequate. Most CNN models neglect its essentiality; for example, SSD [10] directly transmits feature maps from the backbone to the head, while YOLOv3 and RetinaNet [11] simply append several convolutional layers after FPN.

The head is a simple structure that only contains several convolutional layers. It can generate regression and classification predictions, including the coordinates of bounding boxes and the class probabilities.

Most well-known object detectors, which are composed of the above modules, could be split into two-stage detectors and one-stage detectors. The central idea of two-stage detectors is to generate region proposals by the region proposal network, then predict sparse output by detecting each proposal, such as R-CNN [12], Fast R-CNN [13], Faster R-CNN [14], and Mask R-CNN [15]. Compared with two-stage detectors, one-stage ones predict dense output straight from CNNs with the goal of improving detection speed while maintaining comparative performance. YOLOv1 [16], YOLOv2 [17], YOLOv3, YOLOv4 [18], SSD, and RetinaNet are examples of one-stage detectors.

Although these detectors are designed for nature images, their applications in remote sensing have made unprecedented progress. For instance, Yuanxin Ye et al. developed a model with the adaptive feature fusion mechanism based on EfficientDet [19]; the authors of [20] improved YOLOv3 by combining DenseNet with YOLOv3 for multi-scale detection; Ke Li et al. proposed DetectIon in Optical Remote sensing images (DIOR), a large-scale dataset, and compared various detectors in DIOR [21]; Zhenfang Qu et al. designed an auxiliary network with CBAM to improve YOLOv3 [22]; the authors of [23] modified YOLOv4 with MobileNet v2 and depth-wise separable convolution to achieve the tradeoff between detection accuracy and speed; and Yafei Jing et al. introduced the vision transformer and Bi-Directional FPN into YOLOv5s [24]. In remote sensing, multi-scale object detection has made obvious advances by transferring and improving existing detectors. However, it still cannot meet the requirements of practical applications, especially in small object detection.

To address the aforementioned problem, we concentrate on the neck of detectors and carefully design neck attention block (NAB), a simple and flexible module which combines the attention mechanism and the convolutional bottleneck structure to enhance the feature representation capability and promote feature transition from feature maps to dense output. It can extract global information and calibrate the channels of feature maps. It can be inserted straightforwardly after the feature maps generated by the backbone or the path aggregation structure. In addition, we propose a publicly dataset, Vehicle in High

Resolution Aerial Imagery (VHRAI) for small object detection. YOLOv3, YOLOv4-Tiny, and SSD were modified simply with NAB, and the improved models were validated on various datasets. By conducting experiments compared with the original models, we demonstrate that NAB is beneficial to small object detection and multi-scale object detection in remote sensing. In addition, it had excellent generalizability on various datasets and models.

The rest of this paper is organized as follows. In Section 2, we introduce some papers about one-stage detectors, attention mechanisms, and small object detection. Section 3 describes NAB, the improved one-stage detectors with NAB, and VHRAI created for small object detection in detail. Section 4 shows the experiments of the improved models on various datasets. Section 5 discusses NAB and the improved models. Lastly, the conclusion is shown in Section 6.

2. Related Work

2.1. One-Stage Detector

In remote sensing, most applications, such as target tracking, military reconnaissance, and disaster relief, have an increased demand for real-time detection. To balance the accuracy and speed of object detection, we concentrate on the research of one-stage detectors.

One-stage detectors can be divided into anchor-based ones and anchor-free ones. For improving recall rate, anchor-based detectors set pre-defined boxes with different scales and ratios for predictions, such as YOLOv2-v4, SSD, and RetinaNet. SSD appends several layers after VGG16 to produce muti-scale output. Based on YOLOv2, YOLOv3 selects the more powerful Darknet-53 as the backbone and uses the FPN to generate multi-scale predictions. YOLOv4 chooses many measures, including CSPNet [25], CIOU [26], and Mosaic, to modify YOLOv3. For real-time detection, YOLOv4-Tiny obtains an extremely higher speed by decreasing the parameters of YOLOv4. Anchor-free detectors directly predict the boxes without the limitation of anchor boxes, such as CornerNet [27], FCOS [28], and YOLOX [29]. FCOS, based on RetinaNet, takes the location, which falls into any ground-truth box, as a positive sample and adds the center-ness branch to depress low-quality predictions. YOLOX proposes more powerful SimOTA as label assignment. Although anchor-free detectors do not need to search for the hyperparameters of anchor boxes and have less complexity, they have lower precision in detecting remote sensing images whose scale of instances changes enormously. By comparing many detectors, we decide to select YOLOv3 and SSD as our baselines to analyze NAB. In addition, we improved YOLOv4-Tiny with NAB for real-time detection.

2.2. Attention Mechanism

Inspired by human vision, attention mechanisms, which enhance meaningful features and depress noise, have shown remarkable improvement in deep learning. In this paper, we focus on the attention mechanisms about CNNs rather than Scaled Dot-Product Attention in Transformer [30]. This method can be well combined with the convolution operation and has lower computational complexity and a faster convergence rate. It can be divided into channel attention and spatial attention. Channel attention focuses on the importance of different channels, and spatial attention focuses on the importance of different locations. In the past few years, many representative blocks have emerged in attention mechanisms, such as SE [31], ECA [32], CA [33], and CBAM [34]. SE is the paradigm of channel attention which adaptively rescales the channels by utilizing the information of feature maps. Figure 1 shows the structure of the SE block in detail. It obtains global information by GAP (global average pooling) and then utilizes two fully connected layers to produce the response of each channel. Finally, channel-wise multiplication is implemented between the response and the original feature map. ECA thinks the two fully connected layers of SE are unnecessary and adopts one-dimensional convolution to achieve local cross-channel interaction. CBAM combines channel attention and spatial attention to acquire the importance of every channel and location. It concatenates the results of GAP and

GMP (global maximum pooling) to extract more robust information, as shown in Figure 2. CA proposes coordinate attention to calculate the width attention and height attention, respectively. Then, CA implements channel-wise multiplication between them. In remote sensing, AAFM based on CBAM is proposed to create the basic block of EfficientNet. MCA-YOLOv5-Light adopts the MCA attention mechanism to extract more productive information [35].

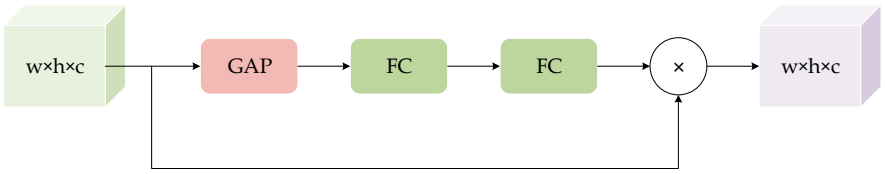

Figure 1. The structure of the SE block. w, h, and c denote the width, height, and channel of a feature map, respectively. 'GAP' is the average-pooling operation along the weight and height axes. 'FC' represents a fully connected layer with an activation function.

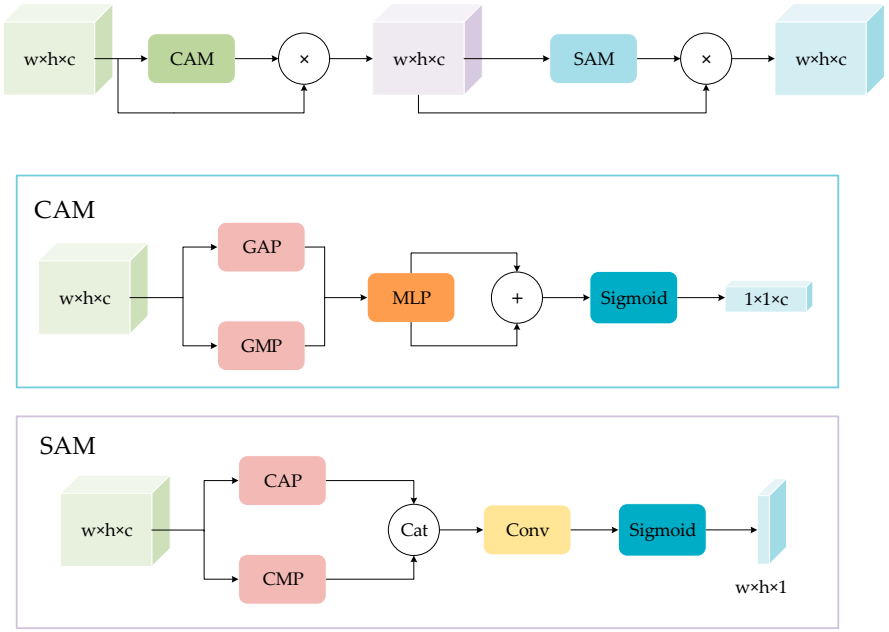

Figure 2. The structure of CBAM. 'CAM' and 'SAM' denote the channel and spatial attention modules, respectively. Similar to 'GAP', 'GMP' is the max-pooling operation along the spatial dimension. Similarly, 'CAP' and 'CMP' are the operations along the channel axis, respectively. The results of 'GAP' and 'GMP' use the identical 'MLP', which is composed of sequential fully connected layers.

Although current studies about attention mechanisms design various architectures in the attention branch, they have a common characteristic that they obtain attention by performing some operations on the feature map and then utilizing the attention to rescale the original feature map. However, NAB, proposed by us, introduces an extra branch to adaptively enhance the information of feature maps, as illustrated in Section 3.1.

2.3. Small Object Detection

In multi-scale detection, detecting small targets which have limited features, diverse distributions, and arbitrary orientations is a big problem. First, there is no uniform definition of small objects. The most universal definition is from MS COCO, which regards objects less than 32 × 32 pixels as small objects [36]. In DOTA [37], an object whose height of the horizontal bounding box ranges from 10 to 50 pixels is defined as a small object. TinyPerson takes an object that ranges from 20 to 32 pixels as a small object [38]. Chen et al.

established a small object dataset whose ratio of the bounding box over the image of all instances was between 0.08% and 0.58% [39]. With the consideration of limited receptive field and down-sampling rate, we selected the above definition of MS COCO for small objects in this paper.

For small object detection in remote sensing, the datasets and related research are inadequate. The significant targets of remote sensing images usually contain almost 20 categories, such as soccer ball fields, vehicles, planes. In these categories, vehicles, ships, and planes, which generally have a large number of small instances. TAS [40], VEDAI [41], and COWC [42] only focus on vehicles; HRSC2016 only contains ships [43]; and UCAS-AOD is concerned with vehicles and planes [44]. These datasets have many instances that do not meet our definition of small objects. DOTA and DIOR contain enormous multi-scale instances, but they do not specialize in annotating small objects.

With respect to object detectors, YOLO-fine, which is based on YOLOv3, increases the resolution of feature maps for detecting small targets [45]. Deconv R-CNN introduces a deconvolution layer to recover more details [46], and SOON constructs a receptive field enhancement module to extract spatial information [47]. Most research ignores the importance of the neck. Our proposed NAB is a flexible module which is used to extract global information and propel the transition of features in the neck.

3. Materials and Methods

3.1. NAB

In the early stages of CNNs, the neck of an object detector, which is usually used to transmit feature maps generated by the backbone to the head, is generally composed of several convolutional layers. With the advent of FPN and PAN, the neck plays another important role in producing multi-scale feature maps that possess strong sematic information by appending top-down and down-top paths. Then, these feature maps are sent into the head via the identical layers. The way of stacking layers in the neck has a large burden for the models with multi-scale output. For example, every output of FPN is connected with 5 convolutional layers in the neck of YOLOv3. This way increases the parameters of YOLOv3 and causes overfitting in the training process, especially for remote sensing datasets that contain inadequate images.

In order to enhance representation capability in the neck, we carefully designed NAB, which combines the channel attention and the convolutional bottleneck structure. It consists of an attention branch, which adopts attention mechanisms to learn where and what to focus on, and a bottleneck branch, which utilizes the convolutional bottleneck structure to refine features and obtain robust feature representation adaptively.

Whereas attention mechanisms contain channel attention and spatial attention, we only utilized channel attention in the attention branch. There is an empirical explanation why we excluded spatial attention: for the dense output of one-stage detectors: Each grid cell predicts the result of the corresponding region in an input image. Every region should be weighted equally. Because the neck is close to the final output, spatial attention would breach this equality and result in bad performance. However, the channels of a grid cell denote different properties, such as the coordinates of a bounding box and the categories. Using channel attention can propel feature representation and convergence.

Inspired by SE, the attention branch adopts GAP to aggregate global information, as illustrated in Equation (1). The input is assumed as X, and $X_{i,j}$ denotes the value of a specific spatial location. Then, the information is forwarded to successive multi-layer perceptrons (MLPs) composed of two fully connected layers. It is notable that the last layer in the branch follows a Sigmoid function to generate factors which are restricted to the range of 0–1.

$$F_{GAP}(X) = \frac{1}{h \times w} \sum_{i=1}^{w} \sum_{j=1}^{h} X_{i,j} \quad (1)$$

The purpose of appending a bottleneck branch is to enhance the adaptive ability of the attention mechanism and to produce a more robust feature map via convolutional layers. We opted to utilize the factors generated by the attention branch to recalibrate the output of the bottleneck branch, which is the most distinctive point compared with traditional attention mechanisms that remodify the original feature maps with the factors. The reasons why this novel method feasible are as follows: The outputs of both branches originate from the identical feature map. This can increase the flexibility of the attention mechanism and make the block refine features adaptively. Our proposed NAB can acquire more robust features and decrease the extra complexity introduced by traditional attention mechanisms.

The structure of NAB is shown in Figure 3. In NAB, the first and second lines denote the attention branch and the bottleneck branch, respectively. The bottleneck branch is composed of 'BB', which contains 3×3 and 1×1 convolution layers. It is notable that each convolutional layer is connected with BN (Batch Normalization) [48] and ReLU. The attention branch is composed of 'GAP' and several 'MLP'. We set 'BB' and 'MLP' to have an identical number, denoted by m. The output feature map has the same size and channel as the input one. Our proposed NAB is an innovation for traditional attention mechanisms which rescale the original feature map. It can decrease the parameters of neck and enhance feature representation ability. If m is 1, then the attention branch and the bottleneck branch can be represented as Equations (2) and (3), respectively. 'F_{fc}' denotes one FC layer with an activation function. 'F_{1c}' and 'F_{3c}' represent 1×1 and 3×3 convolutional layers with Batch Normalization and a ReLU function, respectively. By implementing channel-wise multiplication, the output of NAB can be obtained, as shown in Equation (4). In Section 4, we show the excellent performance of NAB for small object detection and multi-scale object detection on various datasets.

$$F_{\text{attention}}(X) = F_{fc}(F_{fc}(F_{\text{GAP}}(X))) \tag{2}$$

$$F_{\text{bottleneck}}(X) = F_{1c}(F_{3c}(X)) \tag{3}$$

$$F_{\text{NAB}}(X) = F_{\text{attention}}(X) \otimes F_{\text{bottleneck}}(X) \tag{4}$$

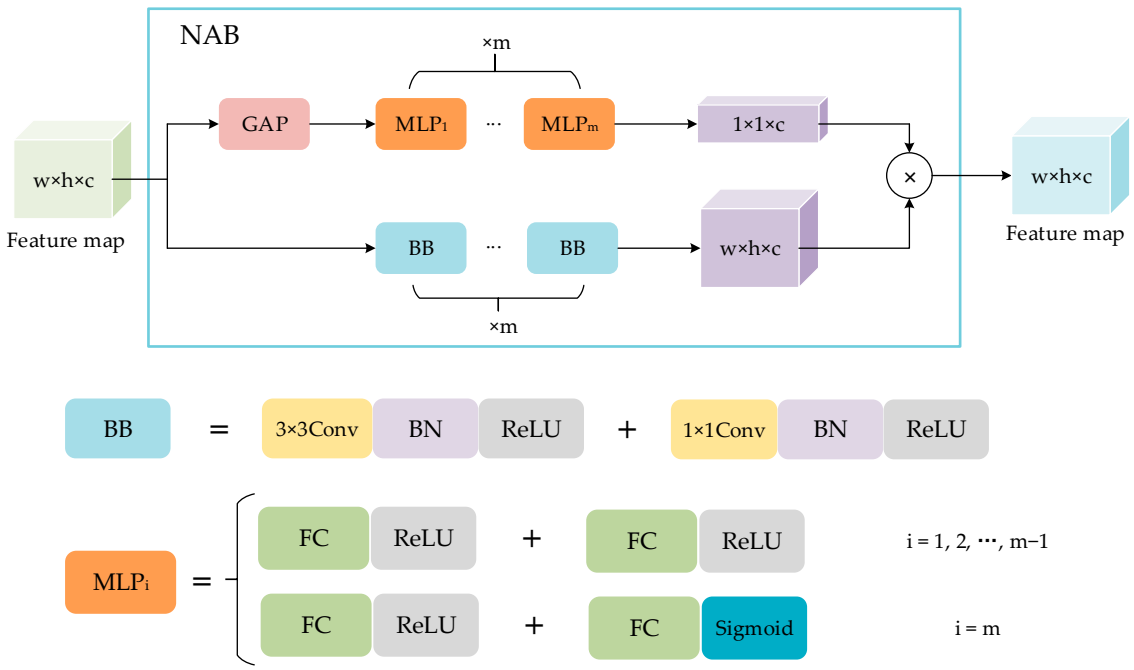

Figure 3. The structure of NAB. 'BB' is the abbreviation of bottleneck block. The number of 'MLP' and 'BB' denoted by m is equal. The second 'FC' in the last MLP uses a Sigmoid activation function to scale the factors into the range of 0–1.

3.2. Improved Models

In remote sensing, the instances, which usually have complicated backgrounds, uneven distributions, and diverse scales, bring enormous computational complexity for object detectors. To balance the accuracy and speed of object detection, we concentrated on the research of one-stage detectors. In addition, NAB, which assigns different attributes to the channels of a feature map, is consistent with the output of one-stage detectors whose every channel denotes a kind of attribute, such as the coordinates of bounding boxes and the probabilities of classes. We selected YOLOv3, YOLOv4-Tiny, and SSD as the improved models from various one-stage detectors.

YOLOv3 is the baseline of many detectors, including YOLOv4, YOLOv5, and YOLOX. It has important values for researching one-stage detectors; therefore, we selected YOLOv3 to verify the effectiveness of NAB. On the basis of YOLOv2, YOLOv3 adopts more powerful Darknet53 as the backbone to enhance the capability of feature extraction and FPN to generate multi-scale output. Due to the way that YOLOv3 transmits semantic information to finer-grained feature maps by the top-down path, it obtains salient performance on small object detection. In the neck, it is notable that YOLOv3 has five sequential 'CBL' blocks before each head. It has an inferior ability in fusing feature maps generated by FPN and Darknet53 and introduces redundant parameters to increase the risk of overfitting. Aiming at achieving higher performance while decreasing computational complexity, the original five 'CBL' were replaced with our proposed NAB and 1×1 'CBL' which was used to reduce the channels of the feature map. Figure 4 depicts the modification in the neck of YOLOv3. In Section 4, we contrast different models and show the highlighted performance of NAB on a variety of datasets.

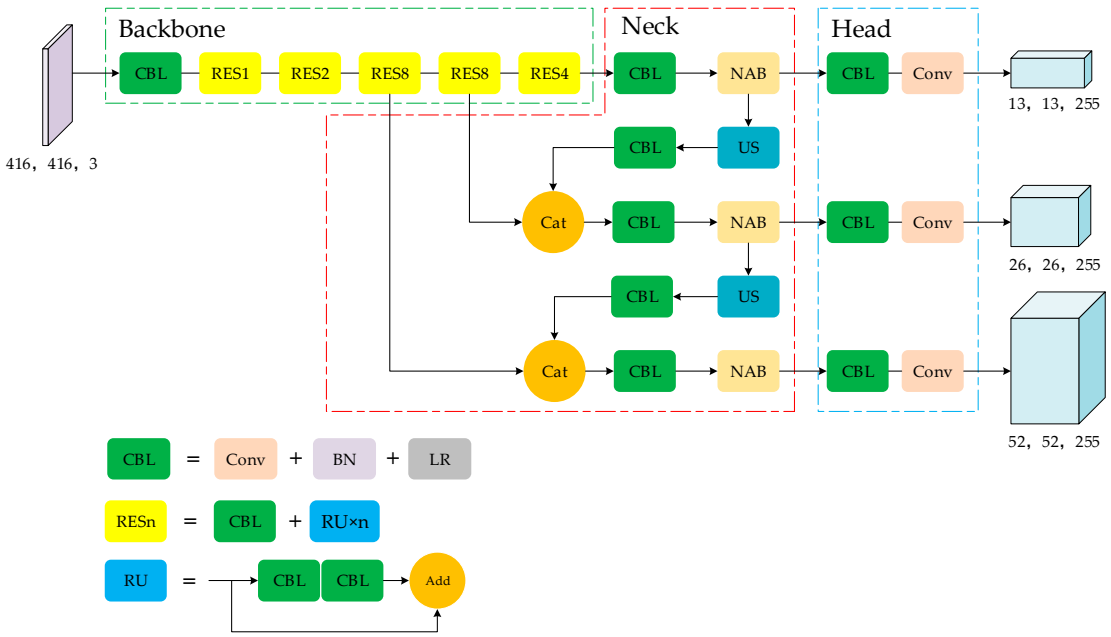

Figure 4. The network of the improved YOLOv3. It can be divided into the backbone, called Darknet53; the neck, which contains FPN and NAB; and the head, which is composed of two convolutional layers. 'Cat' and 'US' denote the operations of concat and up-sampling, respectively. 'LR' is the abbreviation of Leaky ReLU.

SSD is another paradigm of one-stage detectors. The backbone is composed of the truncated VGG16 and several auxiliary convolutional layers. It selects six multi-scale feature maps that are generated by different convolutional blocks of the backbone. Then, these feature maps are transmitted to the corresponding detection layers in the head. Each layer has two convolutional layers, one for predicting the probabilities of classes and the other for predicting the information of bounding boxes. If the input size is 300×300,

then SSD will generate 8732 outputs. Because of the large size of optical remote sensing images, the application of SSD in remote sensing has extremely tremendous computation complexity and low detection efficiency. As a result, we improved SSD for validating the generality of NAB in nature images rather than remote sensing images. In the original SSD, the author introduced 'L2_norm' to scale the feature map of 'Conv_4', which is different from others. Because NAB also has the same function, we concisely removed the 'L2_norm'. We inserted NAB and 'CBL' between the backbone and the head of SSD to enhance the capability of feature representation and facilitate the feature transition, as depicted in Figure 5. Section 4.3 shows the excellent generalizability of NAB in detecting nature images.

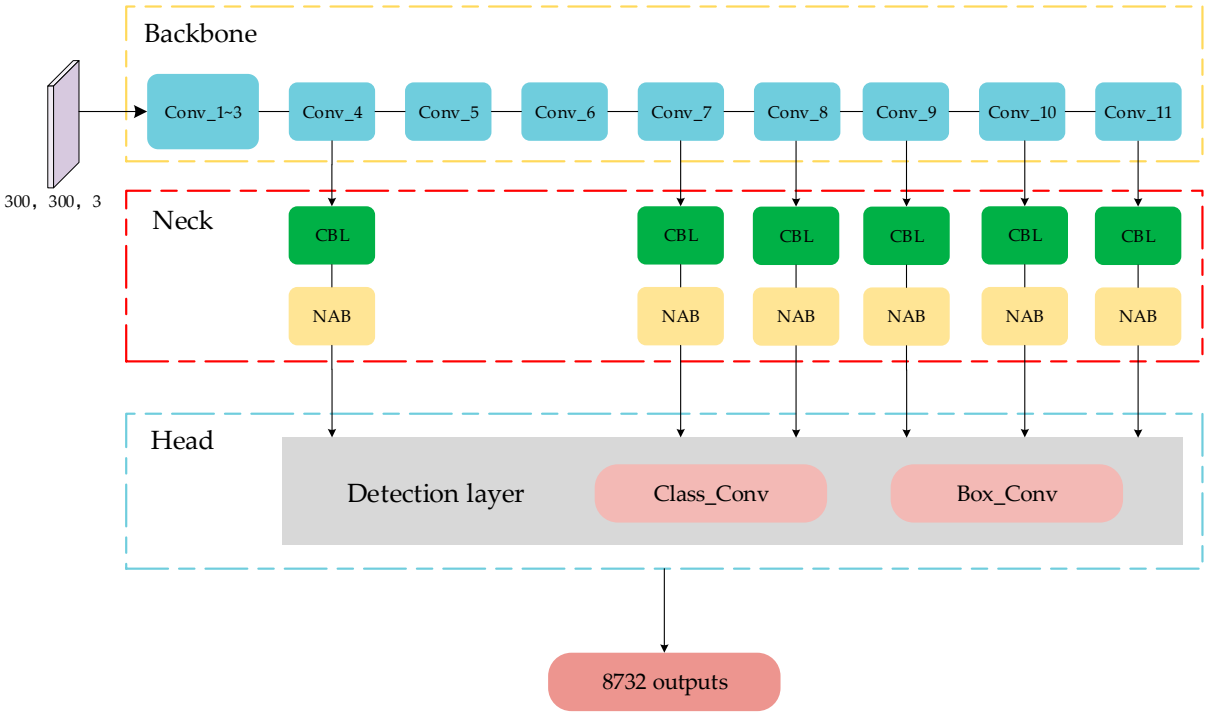

Figure 5. The network of the improved SSD. 'Conv_1~3' is the first three blocks of VGG16.

For real-time detection, YOLOv4-Tiny, which has an excellent balance between accuracy and speed, was modified with NAB. It is a simple version of YOLOv4 and has only about one-tenth of YOLOv4's parameters. YOLOv4-Tiny introduces the idea of CSPnet, which is the largest difference between it and YOLOv3. It only has two outputs for reducing the parameters. We improved it by inserting NAB into the neck, as shown in Figure 6. Because YOLOv4-Tiny has fewer channels than YOLOv3 and SSD, we did not append 'CBL' before NAB. By conducting experiments on various datasets, we found that the improved YOLOv4-Tiny has a more powerful capability in multi-scale object detection than the original one, though it increases computational complexity slightly.

3.3. Datasets

Deep learning is a science driven by data. Thanks to substantial datasets that are available in remote sensing, multi-scale object detection with CNNs has made remarkable progress. However, small object detection remains a challenge. Besides the characteristics of small targets, another reason is the lack of appropriate datasets that specialize in detecting small instances. In order to boost the performance of small object detection, we created Vehicle in High-Resolution Aerial Imagery (VHRAI), a dataset that contains 900 aerial images with 960 × 540 pixels captured at a height of 1000 m for vehicle detection. We utilized LabelImg, an open-source image annotation tool, to annotate instances [49]. Each

object instance was manually labeled by a horizontal bounding box which was composed of the coordinates of the central point, the size of the box, and the category.

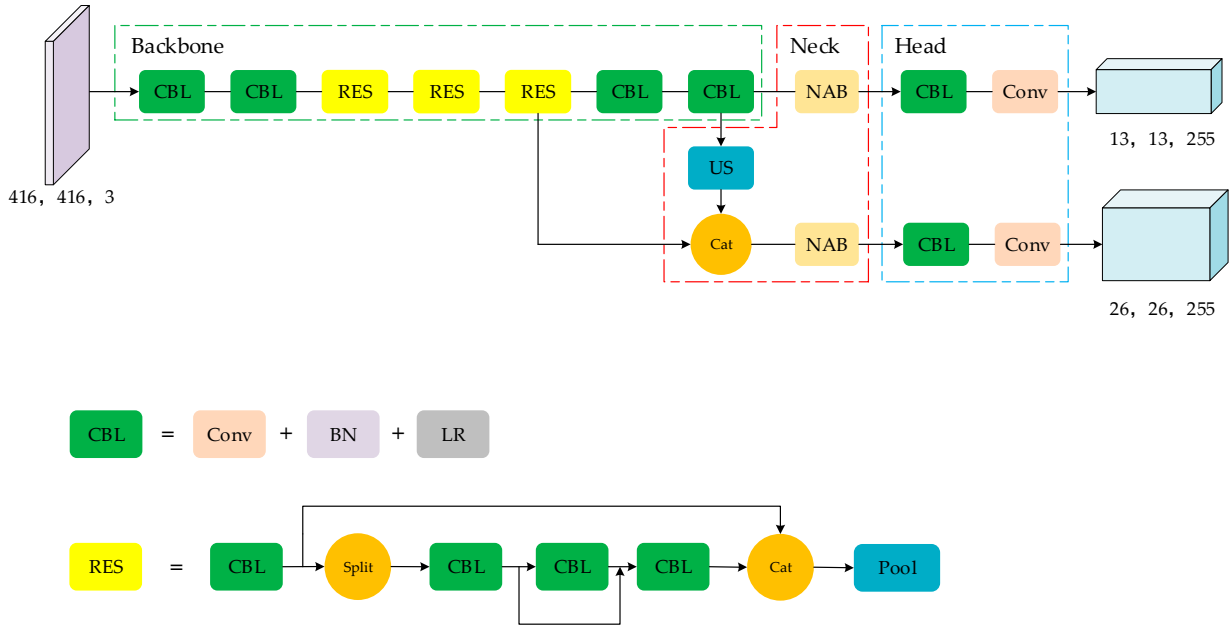

Figure 6. The network of the improved YOLOv4-Tiny. 'Split' is the operation that divides the feature map into two portions along the channel axis. 'Pool' denotes 2 × 2 max-pooling.

Because VHRAI is created for small object detection, we compared it with some well-known datasets which mainly concentrate on researching vehicles and ships, including TAS, UCAS-AOD, HRSC2016, DLR-MVDA [50], COWC, and VEDAI, as listed in Table 1. The average area per instance of DLR-MVDA and VHRAI is far smaller than other datasets. DLR-MVDA and VHRAI are annotated with oriented bounding boxes (OBB) and horizontal bounding boxes (HBB), respectively. Both have important value in object detection. Compared with VEDAI (512), VHRAI has more instances and smaller bounding boxes. However, VHRAI has fewer instances than some large datasets, including UCAS-AOD and COWC. In the future, we will further capture more images to enlarge VHRAI.

Table 1. Comparisons between the proposed VHRAI and several publicly available datasets in remote sensing. VEDAI (512) denotes the version of VEDAI, whose image width is 512. Because the annotations of the testing set in DLR-MVDA are unavailable, we only display the properties of the training set. '#' represents the meaning of 'the number of'.

Datasets	# Categories	# Images	# Instances	Image Width	Average Area per Instance
TAS	1	30	1319	792	805
UCAS-AOD	2	1510	14,597	1280	4888
HRSC2016	1	1070	2976	~1000	56,575
DLR-MVDA	2	10	3505	5616	239
COWC	1	53	32,716	2000~19,000	1024
VEDAI (512)	9	1250	3757	512	3108
VHRAI (ours)	1	900	5589	960	369

Figure 7 shows the characteristics of VHRAI. It has diverse backgrounds, uneven distributions, and tiny scales. In the Earth observation community, VHRAI is a challenging dataset for small object detection.

Figure 7. Examples in VHRAI.

To validate the effect of NAB for multi-scale object detection in remote sensing, we selected TGRS-HRRSD, a public dataset which has 21,761 images and 13 categories [51]. This elaborate dataset achieves an excellent balance between all categories. The average scale per category of TGRS-HRRSD ranges from 41.96 to 276.50 pixels. In addition, aiming at indicating the generalizability of NAB, we conducted experiments on PASCAL VOC [52], a commonly used natural scene dataset. The entire results are displayed in the next section.

4. Results

4.1. Evaluation Criteria

The output of an object detector can be divided into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP and FP denote a positive sample that is classified correctly and incorrectly, respectively. TN and FN represent a negative sample that is classified correctly and incorrectly, respectively. Through analyzing these categories, we can obtain Precision, which illustrates the proportion of TP in all positive samples, and Recall, which indicates the proportion of TP in all positive ground-truth samples, depicted in Equations (5) and (6). These two indicators have some limitations as evaluation criteria. Confidence is the threshold that estimates a sample is positive or negative. Different Confidence can generate different Precision and Recall. Precision increases and Recall decreases in general as Confidence gradually increases.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

We can acquire the Precision/Recall Curve by setting a different Confidence. Average Precision (AP) denotes the area under Precision/Recall Curve for a class. The mean

Average Precision (mAP) is the mean value of every class's AP. The mAP is a significant evaluation criterion in object detection. In this paper, we considered Precision and Recall for a comprehensive comparison. AP_s proposed in MS COCO was also adopted to show the performance of detecting small objects. In addition, the number of parameters was used to evaluate computational complexity and detection speed.

4.2. VHRAI

VHRAI, whose average size of instances is 19.22 × 19.19 pixels, is a valuable dataset for small object detection. On this dataset, we validated the effectiveness of NAB by comparing the improved YOLOv3 and YOLOv4-Tiny with the original ones. We also compared traditional methods that contained SE and CBAM with NAB to reveal the importance of the bottleneck branch in NAB, as shown in Figure 8. Furthermore, we analyzed the influence of 'm', a hyperparameter in NAB. It is notable that these improvements were adopted in all multi-scale paths.

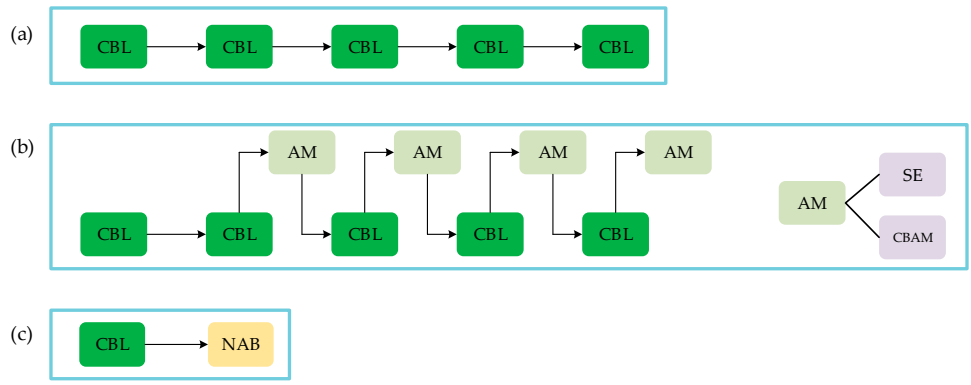

Figure 8. (**a**) The original 'CBL5'in the neck. (**b**) The improved 'CBL5'with the attention mechanism in the neck. (**c**) NAB in the neck. The structure of 'CBL' is shown in Figure 4. 'AM', which is the abbreviation of attention mechanism, could be SE or CBAM.

All experiment results on VHRAI are listed in Table 2. YOLOv3-NAB (m = 1) obtained the best AP 90.29% among all models, surpassing YOLOv3 by 1.94%. The accuracy and recall of YOLOv3-NAB were also better than the original model. At the same time, YOLOv3-NAB (m = 1) reduced parameters by almost 11% compared with YOLOv3. We also compared the loss and mAP curves of YOLOv3-NAB (m = 1) and YOLOv3 to acquire a more robust conclusion, as shown in Figure 9. With respect to traditional attention mechanisms, YOLOv3-SE had a slightly poorer performance than YOLOv3, and YOLOv3-CBAM exceeded YOLOv3 by 0.74%. However, YOLOv3-NAB (m = 1), which improved the attention mechanism by appending a bottleneck branch, achieved more salient performance and had fewer parameters compared with YOLOv3-CBAM and YOLOv3-SE. The reason why YOLOv3-NAB (m = 1) had fewer parameters is that traditional attention mechanisms only can rescale the feature map generated by the layer and cannot serve as an independent module. With this limitation, YOLOv3-SE had more convolutional layers than YOLOv3-NAB (m = 1). These results clearly show the effectiveness of NAB, which is a better way to utilize the attention mechanism in the neck. In addition, when we set 'm' = 2, YOLOv3-NAB (m = 2) obtained worse AP and had more parameters than YOLOv3-NAB (m = 1). This may be attributed to the fact that more parameters increase the risk of overfitting with the limitation of inadequate images. In the next experiments, the default value of 'm' in NAB was 1.

Table 2. Detection results on VHRAI. For a fair comparison, the models based on YOLOv3 had the same configuration. The models based on YOLOv4-Tiny also had the same configuration. YOLOv3-SE and YOLOv3-CBAM adopted the improved 'CBL5' with attention mechanisms in Figure 8b.

Model	# Parameters	Precision (%)	Recall (%)	AP (%)	AP_s (%)
YOLOv3	61.52 M	83.73	83.87	88.35	41.9
YOLOv3-SE	63.24 M	84.44	83.5	88.15	39.5
YOLOv3-CBAM	63.24 M	85.88	83.76	89.09	41.4
YOLOv3-NAB (m = 1)	54.81 M	86.62	84.36	90.29	42.9
YOLOv3-NAB (m = 2)	61.87 M	81.01	87.45	89.16	41.9
YOLOv4-Tiny	5.87 M	71.35	58.77	63.99	20.6
YOLOv4-Tiny-NAB	7.05 M	72.05	60.39	65.82	21.6

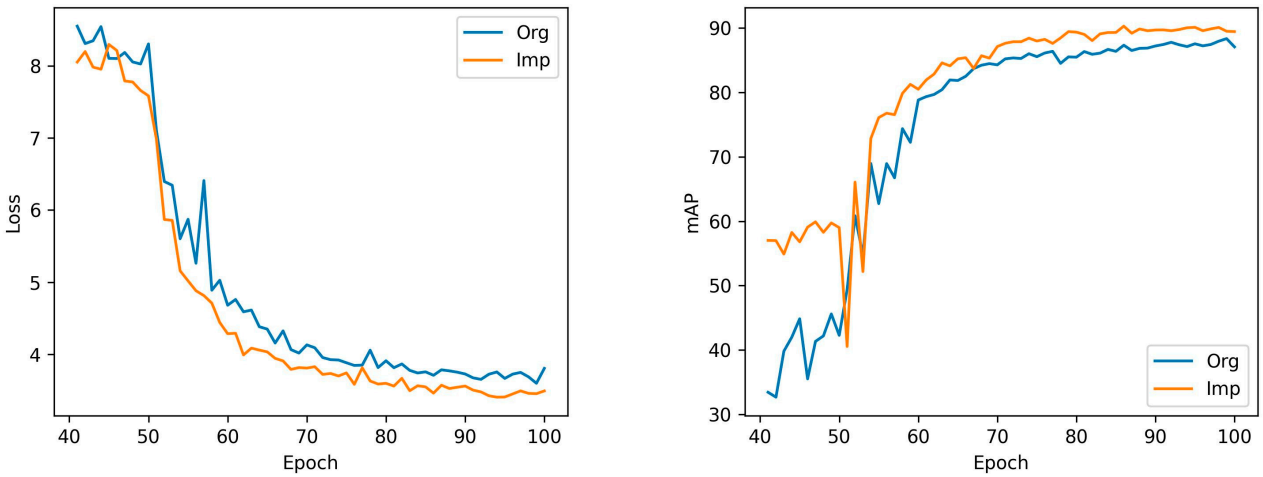

Figure 9. The loss and mAP curves of YOLOv3 and YOLOv3-NAB (m = 1). 'Org' and 'Imp' denote YOLOv3 and YOLOv3-NAB (m = 1), respectively.

For real-time detection, we conducted experiments on YOLOv4-Tiny. YOLOv4-Tiny-NAB achieved better precision, recall, and AP, exceeding the original model by 0.7%, 1.62% and 1.83%, respectively, despite having slightly more parameters. Furthermore, due to YOLOv4-Tiny, which cut an important path for small object detection, we found that YOLOv4-Tiny-NAB had a large gap in performance compared with YOLOv3-NAB (m = 1). It is notable that YOLOv4-Tiny-NAB had an extremely fast speed in detection.

In addition, we compared the AP_s of the above models, which is used to evaluate the performance for small object detection precisely in MS COCO. Undoubtedly, YOLOv3-NAB (m = 1) obtained the best AP_s, outperforming YOLOv3-SE and YOLOv3-CBAM by 3.4% and 1.5%, respectively. In addition, YOLOv4-Tiny-NAB was better than YOLOv4-Tiny. These experiments on VHRAI apparently demonstrate the effectiveness of NAB in small object detection. Different from traditional attention mechanisms, we introduced an extra branch to enhance the ability of adaptively extracting features rather than focusing on designing a more complicated attention branch. Furthermore, NAB can be inserted into a model flexibly as an independent structure, similar to the above models.

4.3. TGRS-HRRSD

Although we proved the effectiveness of NAB in small object detection, which is a crucial part of multi-scale detection, it is necessary to conduct experiments on a multi-scale dataset to acquire a reliable conclusion. TGRS-HRRSD is a large dataset for multi-scale object detection. It has 13 categories, and the average scale per category ranges from 41.96 pixels to 276.50 pixels. We selected TGRS-HRRSD as the dataset and com-

pared YOLOv3-NAB, YOLOv3-SE, YOLOv3-CBAM, and YOLOv4-Tiny-NAB with the original models.

Table 3 shows the detection results. YOLOv3-NAB, which had fewer parameters than YOLOv3, scored 92.16% mAP, surpassing YOLOv3 by 1.06%. With respect to traditional attention mechanisms, YOLOv3-SE had an inferior performance than YOLOv3-NAB, and YOLOv3-CBAM was comparable with YOLOv3-NAB, but its parameters increased by 13%. Compared with YOLOv4-Tiny, YOLOv4-Tiny-NAB, which was improved with NAB in the neck, obtained a remarkable performance that exceeded the original model by 3.72% mAP. It also outperformed in all categories. Its mAP was even close to YOLOv3, though it only had about one-tenth of YOLOv3's parameters. These experiments clearly illustrate that NAB can obtain robust feature representation and is helpful for multi-scale object detection as a flexible module.

Table 3. Detection results on TGRS-HRRSD.

Model	Ship	Bridge	Ground Track Field	Storage Tank	Basketball Court	Tennis Court	Airplane	Baseball Diamond	Harbor	Vehicle	Crossroad	T Junction	Parking Lot	mAP (%)
YOLOv3	92.65	92.04	98.40	93.99	83.17	96.06	99.57	93.05	95.02	92.69	93.92	83.62	70.09	91.10
YOLOv3-SE	94.37	92.72	98.31	94.47	84.05	95.81	98.73	93.62	92.79	96.91	92.31	82.86	70.67	91.35
YOLOv3-CBAM	94.63	92.78	98.71	96.89	82.35	95.12	99.54	93.63	97.27	97.02	92.51	84.72	72.03	92.09
YOLOv3-NAB	94.59	93.33	98.33	96.12	82.84	95.83	99.02	93.34	96.79	97.05	94.08	85.03	71.79	92.16
YOLOv4-Tiny	86.34	73.18	92.31	97.20	69.60	93.53	98.88	89.90	84.36	90.27	87.13	68.85	53.36	83.44
YOLOv4-Tiny-NAB	89.72	85.31	95.89	97.28	71.30	93.61	98.94	91.61	92.11	93.43	89.98	73.15	60.79	87.16

4.4. PASCAL VOC

NAB had excellent performance in multi-scale remote sensing images, and we speculate that it is not limited in remote sensing. The experiments on PASCAL VOC, a well-known dataset that contains 21504 nature images, were conducted to validate the generalizability of NAB. All models were trained on the union of VOC2007 and VOC2012 trainval, and they were evaluated with the VOC2007 test. The detection results are shown in Table 4. The improved models, including YOLOv3-NAB, YOLOv4-Tiny-NAB, and SSD-NAB, acquired a better performance than the original models, surpassing them by 0.88%, 1.98%, and 0.82% mAP, respectively. Compared with traditional mechanisms, YOLOv3-NAB outperformed YOLOv3-SE and YOLOv3-CBAM by 0.86% and 0.57%, respectively, while decreasing 13% parameters. Consequently, we confirm that NAB can be generalized to natural scenes and applied to various one-stage detectors.

Table 4. Detection results on PASCAL VOC. We chose 10 categories at random to show the comparisons of their AP (%). We retrained YOLOv3, YOLOv4-Tiny, and SSD for a fair comparison using the same configuration of the improved models.

Model	Aero	Bike	Bird	Bottle	Car	Cow	Dog	Horse	Sofa	Train	mAP (%)
YOLOv3	88.84	85.88	80.11	63.78	90.90	84.40	86.62	86.99	73.91	88.86	80.47
YOLOv3-SE	89.28	86.81	81.81	63.21	90.89	83.38	86.82	86.46	78.58	89.80	80.49
YOLOv3-CBAM	89.18	87.26	82.68	62.37	91.32	84.31	85.54	89.97	80.94	89.86	80.78
YOLOv3-NAB	89.65	87.78	81.74	65.97	91.02	86.89	86.38	87.18	71.55	88.70	81.35
YOLOv4-Tiny	84.06	85.29	74.35	62.55	90.54	81.41	77.86	86.55	73.10	84.39	77.08
YOLOv4-Tiny-NAB	87.21	87.09	76.97	66.72	91.77	80.75	79.52	87.67	73.18	86.75	79.06
SSD	77.81	84.93	75.35	42.32	86.50	77.35	86.99	88.60	73.52	84.93	75.8
SSD-NAB	79.75	85.06	77.55	45.79	85.61	77.15	85.67	87.80	74.46	86.66	76.62

5. Discussion

In this paper, we presented NAB, an architectural module designed to enhance the capability of feature representation and promote feature transition in the neck by combining the attention mechanism and the convolutional bottleneck structure. Since the output of NAB has the same dimensions as the input, it is simple and flexible to utilize in the neck of one-stage detectors. In addition, VHRAI, whose instances have an extremely small size,

was created to address the challenge of small object detection, which is a crucial part of multi-scale object detection in remote sensing.

We improved several classic one-stage detectors with NAB and conducted many experiments on various datasets. YOLOv3-NAB, YOLOv4-Tiny-NAB, and SSD-NAB achieved a better performance than the original models. Figure 10 shows some detection examples of YOLOv3-NAB on different datasets. VHRAI is a diverse and challenging dataset whose average area per instance is 369 pixels. It perfectly meets our definition of small objects and is smaller than other famous datasets, such as VEDAI (512), HRSC2016, UCAS-AOD, and TAS. The results of YOLOv3-NAB and YOLOv4-Tiny-NAB on VHRAI clearly illustrate the effectiveness of NAB for small object detection. Furthermore, we compared NAB with SE and CBAM in the neck. NAB, which appends a convolutional bottleneck branch, showed better flexibility and performance while decreasing the model's parameters. The experiments on TGRS-HRRSD demonstrate the capability of NAB for multi-scale detection in remote sensing. YOLOv4-Tiny had poor feature representation with the limitation of parameters. However, the result that YOLOv4-Tiny-NAB exceeded the original model by 3.72% mAP clearly verifies NAB can enhance the capability of feature representation and propel feature transition. These improved models also obtained better detection precision on PASCAL VOC. This fact indicates that NAB is not limited in remote sensing and can be generalized to detect nature images.

Whereas NAB brings promising results in remote sensing and natural scenes, some issues still remain and call for further research. NAB is restricted in the neck of a model. Whether it can replace traditional attention mechanisms in the backbone or not remains unknown. In addition, there are some similarities in the neck of one-stage and two-stage detectors. In the future, we will further focus on the backbone and improve two-stage detectors with NAB for object detection.

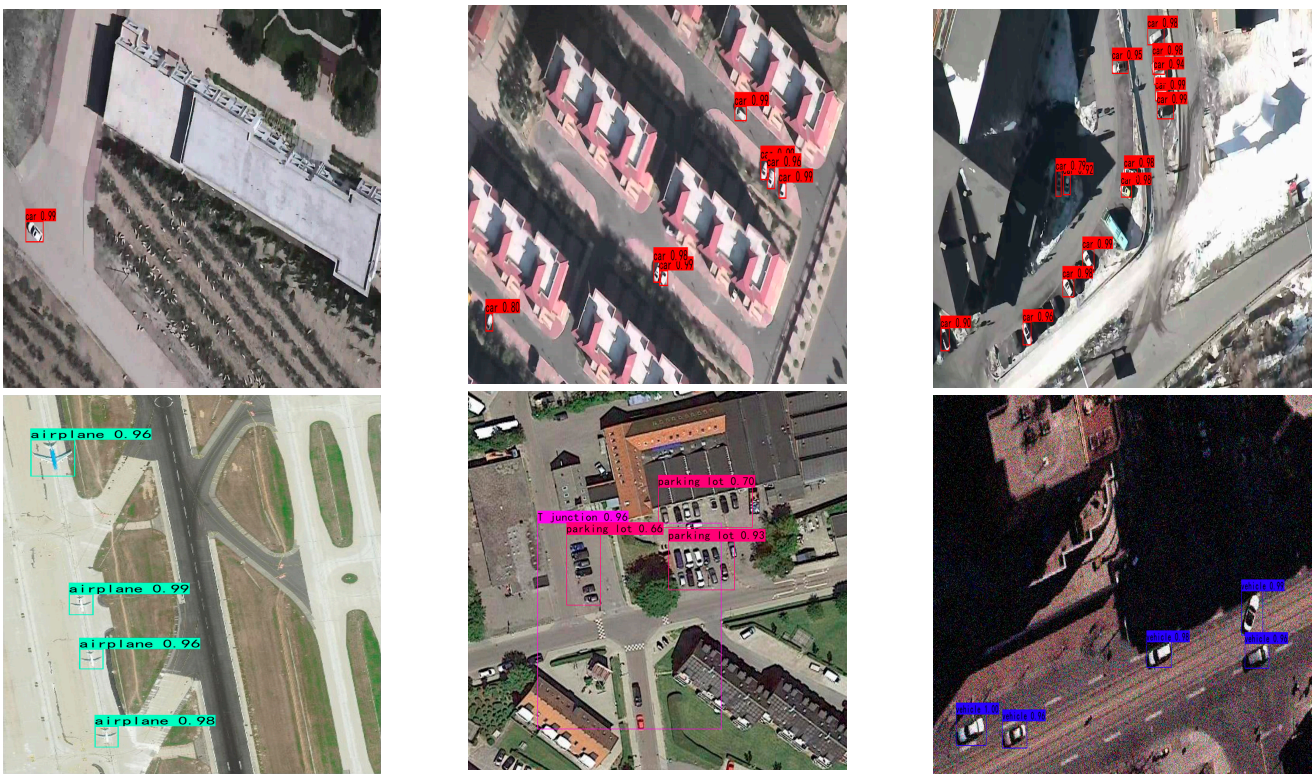

Figure 10. *Cont.*

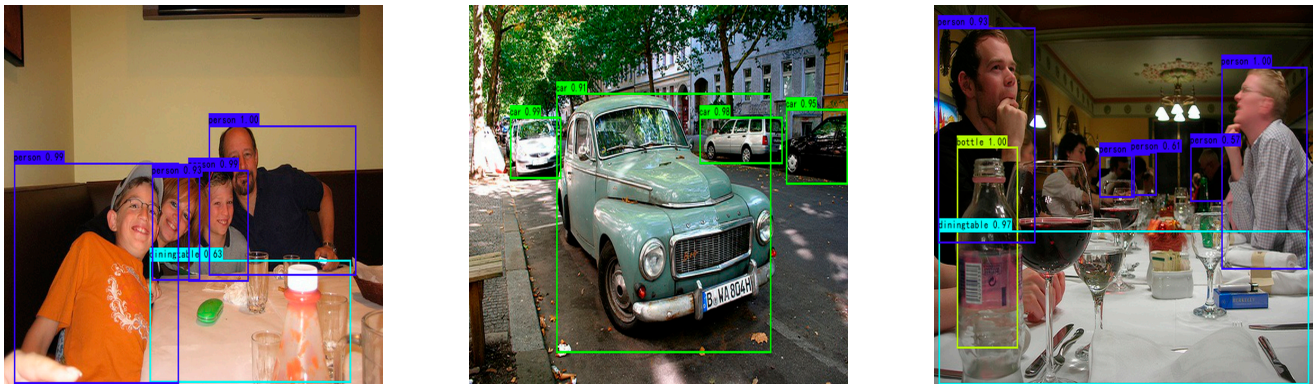

Figure 10. Detection examples of YOLOv3-NAB on various datasets.

6. Conclusions

In this paper, we designed a simple and flexible module for the neck of a model, called NAB. Unlike traditional attention mechanisms which focus on designing a more complicated attention branch, NAB appends a convolutional bottleneck branch with the attention branch for enhancing feature representation capability and promoting feature transition. In addition, VHRAI, a challenging dataset whose instances have an extremely small size, was proposed for small object detection. The improved models, including YOLOv3-NAB, YOLOv4-Tiny-NAB, and SSD-NAB, achieved excellent performance on various datasets, which clearly proves the effectiveness and generalizability of NAB in small object detection and multi-scale object detection. In the future, we will focus on improving the backbone of a model and two-stage detectors on the basis of NAB.

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

Funding: This research was funded by the Scientific Research Business Fee Fund of Heilongjiang Provincial Scientific Research Institutes, Research on Key Technologies of Wide Area Forest and Grass Fire Aerial Monitoring and Early Warning (No. CZKYF2020B009).

Data Availability Statement: All data used during the study have been uploaded at https://github.com/youjiaowuya/NAB (accessed on 1 June 2020).

Acknowledgments: This research was supported by Foundation items: The National Key Research and Development Program of China.

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

References

1. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
2. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
4. Szegedy, C.; Wei, L.; Yangqing, J.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
6. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.

7. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
8. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
9. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot MultiBox detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland; pp. 21–37.
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
17. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
18. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
19. Ye, Y.; Ren, X.; Zhu, B.; Tang, T.; Tan, X.; Gui, Y.; Yao, Q. An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images. *Remote Sens.* **2022**, *14*, 516. [CrossRef]
20. Xu, D.; Wu, Y. Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. *Sensors* **2020**, *20*, 4276. [CrossRef]
21. 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]
22. Qu, Z.; Zhu, F.; Qi, C. Remote Sensing Image Target Detection: Improvement of the YOLOv3 Model with Auxiliary Networks. *Remote Sens.* **2021**, *13*, 3908. [CrossRef]
23. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight Underwater Object Detection Based on YOLO v4 and Multi-Scale Attentional Feature Fusion. *Remote Sens.* **2021**, *13*, 4706. [CrossRef]
24. Jing, Y.; Ren, Y.; Liu, Y.; Wang, D.; Yu, L. Automatic Extraction of Damaged Houses by Earthquake Based on Improved YOLOv5: A Case Study in Yangbi. *Remote Sens.* **2022**, *14*, 382. [CrossRef]
25. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1571–1580.
26. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI 2020 Conference, New York, NY, USA, 7–12 February 2020.
27. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
28. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
29. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
30. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1312.4400.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; 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, 13–19 June 2020; pp. 11531–11539.
33. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
34. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
35. Sun, C.; Zhang, S.; Qu, P.; Wu, X.; Feng, P.; Tao, Z.; Zhang, J.; Wang, Y. MCA-YOLOV5-Light: A Faster, Stronger and Lighter Algorithm for Helmet-Wearing Detection. *Appl. Sci.* **2022**, *12*, 9697. [CrossRef]

36. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
37. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
38. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale match for tiny person detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1246–1254.
39. Chen, C.; Liu, M.-Y.; Tuzel, O.; Xiao, J. R-CNN for small object detection. In Proceedings of the ACCV 2016—Asian Conference on Computer Vision, Taipei, Taiwan, 21–23 November 2016.
40. Heitz, G.; Koller, D. Learning spatial context: Using stuff to find things. In Proceedings of the ECCV 2008—10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008.
41. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
42. Mundhenk, T.N.; Konjevod, G.; Sakla, W.A.; Boakye, K. A Large contextual dataset for classification, detection and counting of cars with deep learning. In Proceedings of the ECCV—Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
43. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
44. 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.
45. Pham, M.-T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2501. [CrossRef]
46. Zhang, W.; Wang, S.-P.; Thachan, S.; Chen, J.; Qian, Y.-t. Deconv R-CNN for small object detection on remote sensing images. In Proceedings of the IGARSS—IEEE International Geoscience Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2483–2486.
47. Qin, H.; Li, Y.; Lei, J.; Xie, W.; Wang, Z. A Specially Optimized One-Stage Network for Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 401–405. [CrossRef]
48. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37, Lille, France, 6–11 July 2015; pp. 448–456.
49. Tzutalin. LabelImg. Git Code. 2015. Available online: https://github.com/tzutalin/labelImg (accessed on 10 September 2020).
50. Liu, K.; Mattyus, G. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942. [CrossRef]
51. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
52. Everingham, M.; Eslami, S.M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]