Improved Dual Attention for Anchor-Free Object Detection

Ye Xiang 1, Boxuan Zhao 1, Kuan Zhao 1, Lifang Wu 1 and Xiangdong Wang 2,*

1 Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; xiangye@bjut.edu.cn (Y.X.); zhaoboxuan@email.bjut.edu.cn (B.Z.); s201861077@email.bjut.edu.cn (K.Z.); lfwu@bjut.edu.cn (L.W.)
2 School of Physical Education, Jimei University, Xiamen 361021, China
* Correspondence: wangxiangdong@jmu.edu.cn

Abstract: In anchor-free object detection, the center regions of bounding boxes are often highly weighted to enhance detection quality. However, the central area may become less significant in some situations. In this paper, we propose a novel dual attention-based approach for the adaptive weight assignment within bounding boxes. The proposed improved dual attention mechanism allows us to thoroughly untie spatial and channel attention and resolve the confusion issue, thus it becomes easier to obtain the proper attention weights. Specifically, we build an end-to-end network consisting of backbone, feature pyramid, adaptive weight assignment based on dual attention, regression, and classification. In the adaptive weight assignment module based on dual attention, a parallel framework with the depthwise convolution for spatial attention and the 1D convolution for channel attention is applied. The depthwise convolution, instead of standard convolution, helps prevent the interference between spatial and channel attention. The 1D convolution, instead of fully connected layer, is experimentally proved to be both efficient and effective. With the adaptive and proper attention, the correctness of object detection can be further improved. On public MS-COCO dataset, our approach obtains an average precision of 52.7%, achieving a great increment compared with other anchor-free object detectors.

Keywords: anchor-free object detection; dual attention; adaptive weight assignment

1. Introduction

Object detection algorithms have been considerably improved based on convolutional neural networks (CNNs). As a fundamental research in the field of computer vision, object detection can be utilized in various visual tasks, such as facial detection [1], object tracking [2], and so on. Currently there are still some challenges remaining in object detection, thus attracting much attention from researchers.

Deep learning based methods have become the mainstream in object detection [3,4], and existing correlative algorithms can be divided into two aspects: anchor-based detectors and anchor-free detectors. The anchor-based detectors [5–7] have developed for a long time and obtained great performance. Anchors, a set of predefined boxes with specific sizes and aspect ratios, are essential in anchor-based detectors, such as Faster R-CNN [8], SSD [9], and YOLOv4 [7]. The anchors are to be regressed to correct areas with the help of labels during training, which will reduce the difficulty of bounding boxes prediction. However, the hyper-parameters, e.g., sizes, aspect ratios, and the number of anchors, may have a great influence on the detection results and have to be designed carefully according to the dataset and the algorithm through adequate experiments.

Anchor-free detectors [10–12] do not involve the hyper-parameters setting related to anchors and aim to predict bounding boxes directly. In recent years, much attention has been drawn on anchor-free detectors, meanwhile a problem of low-quality detections often existing in results is also found. To tackle it, FCOS [13] and CornerNet [14] both propose to assign high weights to central regions and low weights to marginal regions within...
bounding boxes, as shown in Figure 1a. The low-quality detections that usually occur at locations far from object center can then be suppressed. Furthermore, a keypoint-based detector named CenterNet [15] is proposed. CenterNet predicts the center keypoint by gathering the max summed values along horizontal and vertical directions. It allows the center keypoints to perceive visual patterns within objects, and the correctness of bounding boxes can be identified consequently. These algorithms explore the weight assignment strategy and believe points near the center should be given more weights to produce high-quality detections. However, the central regions may include nothing of target objects or be shielded in some situations [16,17]. Hence we are to pursue an adaptive weight assignment strategy to obtain the more reasonable confidence scores automatically for predicted bounding boxes, as shown in Figure 1b.

![Input Image](a) ![Weight Assignment Result](b)

**Figure 1.** The visualization of weight assignment results for FCOS (a) and our algorithm (b).

The attention mechanism [18,19] offers a good option for learning weights adaptively, and has been widely utilized in various tasks such as image classification, detection and segmentation [20,21]. There are commonly two types of attention, i.e., the spatial attention and channel attention. Some algorithms [22] concentrate on improving either one of them to boost the performance, whereas more recent works [23–25] are inclined to explore the dual attention to model the richer interdependencies of features. To implement the dual attention mechanism and connect it with deep neural networks, two ways are often exploited: (i) estimate the channel attention and spatial attention successively [25] and (ii) estimate the dual attention in parallel and then fuse them [23,24]. For (i), channel attention is followed with spatial attention, which means channel attention will inevitably interfere spatial attention. Conversely, the parallel way is more likely to learn the effective spatial and channel attention weights directly from original deep features. For (ii), in spatial attention mechanism within a parallel framework, the standard convolutions are generally used, which actually involve interaction in both the spatial and channel dimensions. As a result, the spatial attention weights are confused with the channel attention weights.

In view of these deficiencies, we propose an improved dual attention mechanism and apply it for the adaptive weight assignment in anchor-free object detection. Firstly, we adopt the parallel way to implement the dual attention mechanism. In spatial attention, the depthwise convolution rather than standard convolution is used to untie the spatial and channel attention thoroughly, avoiding the confusion issue. In channel attention, the 1D convolutional layer performing local cross-channel interaction strategy, instead of fully connected layer, is used for higher computational efficiency and better performance. Through improved dual attention mechanism, proper confidence scores for object classification can be obtained adaptively. Since the confidence scores indicate the probability of being high-quality bounding boxes, the correctness of object detection can then be enhanced through NMS [26]. Experiments are conducted on the MS-COCO dataset [27], a challenging benchmark for large-scale object detection. It demonstrates that the proposed algorithm can achieve competitive performance, in contrast with other anchor-free object detectors.

The contributions of this paper are summarized as follows:

- We introduce the adaptive weight assignment based on dual attention mechanism into anchor-free object detector. It can obtain the adaptive distribution of classification confidence scores automatically.
• We improve the dual attention mechanism designed for other tasks and apply it to object detection. The improved dual attention mechanism can prevent the confusion between spatial and channel attention, and is both efficient and effective.
• The experimental results on public MS-COCO dataset demonstrate that the proposed algorithm can achieve the state-of-the-art performance comparing with other anchor-free object detectors.

2. Related Work
There are commonly two kinds of object detection algorithms: anchor-based detectors and anchor-free detectors. Many of them improve performance based on the attention.
Anchor-based detectors. The anchor boxes, which are placed densely on feature maps to adjust and refine prediction results for obtaining the final bounding boxes, are widely used in anchor-based detectors. Faster R-CNN [8] introduced the region proposal network (RPN) with anchors for regions of interest (RoIs) generation. Mask R-CNN [28] appended a mask generation branch to Faster R-CNN [8] for instance segmentation. Cascade R-CNN [29] learned a multi-stage detector by increasing the IoU threshold for more accurate prediction. YOLOv2, YOLOv3, and YOLOv4 [5-7] have been the representative real-time object detection algorithms which also involve the anchors. Another real-time detector SSD [9] assigned the dense anchors on feature maps for object detection with various sizes. RetinaNet [30], which involved a great number of anchors, adopted the focal loss for the problem of imbalance between positive and negative samples. At present, many anchor-based detectors show competitive performance.

Anchor-free detectors. Anchor-free detectors do not require the dense assignment of anchors. CornerNet [14] generated the bounding box through a pair of corners. Based on CornerNet, CenterNet [15] added a center prediction branch to filter the false boxes. CornerNet-Lite [31] combined CornerNet-Saccade and CornerNet-Squeeze to improve both the efficiency and accuracy. The four extreme points of objects were utilized and predicted in ExtremeNet [10]. Zhu et al. [12] predicted the center points and the sizes of bounding boxes for object detection. MatrixNets [32] provided a new deep architecture which can handle objects with different sizes and aspect ratios. ResPoints [11] regarded the object as a fixed number of points that were combined in specific manners for predicted box. YOLOv1 [33] divided the whole image with $S \times S$ grids and regressed the boxes through grid cells containing the centers of objects. DenseBox [34] and UnitBox [35] generated results with the distances from positive points to boundaries of boxes. For positive points, PPDet [17] provided a new definition to implement the relaxed labeling strategy aimed to reduce the label noise in anchor-free detectors. In FSAF [36], an anchor-free branch was attached to RetinaNet [30] for online feature level selection. FCOS [13] equipped with GroupNorm [37] and GIoU loss [38], innovatively introduced the centerness strategy, which assigned the central region with the highest weight to improve the quality of bounding boxes. DDBNet [16] dived deeper into box regressions of center key-points and takes care of semantic consistencies of center key-points.

Attention weight aware. The attention module [18,19] has been widely used in object detection. For the attention in channel dimension, SENet [39] employed the Squeeze and Excitation to learn the weights of channels. SKNet [40] exploited two branches with different kernel sizes to learn the channel weights. ECANet [22] improved the performance of SENet with fewer parameters, which replaced the fully-connected layers with 1D convolutional layers that have dynamic kernel sizes. ResNeSt [41] introduced the split-attention block with the combination of group convolution and channel attention mechanism. The attention modules in these algorithms can be extensively utilized in visual classification, detection, tracking, and segmentation tasks by combining with the backbone network, e.g., ResNet [42]. In this paper, we will design a dual attention module, and add it into object detection network to learn the spatial and channel weights for classification confidence scores.
3. The Proposed Approach

3.1. Overview

We firstly introduce the whole framework of the proposed object detection approach, as shown in Figure 2. There are five components in total: backbone, feature pyramid, adaptive weight assignment based on dual attention, regression, and classification. The backbone is to extract the deep features of objects, and commonly adopts the ResNet-50 [42] or ResNet-101 [42]. The feature pyramid network [43] is utilized to integrate the features of high levels and low levels. In both anchor-based detectors (e.g., RetinaNet [30] and Faster R-CNN [8]) and anchor-free detectors (e.g., FSAF [36] and FCOS [13]), it has been proved that the feature pyramid can enhance the detection performance by a large margin. We employ five levels \{P_3, P_4, P_5, P_6, P_7\} in the feature pyramid with strides of [8, 16, 32, 64, 128], separately. The adaptive weight assignment module based on dual attention aims at weighting confidence scores adaptively for bounding boxes, and is the major contribution of this paper. There are two parallel branches in this module, i.e., the spatial weight assignment and the channel weight assignment. Lastly, the regression subnet and classification subnet follow with feature pyramid and adaptive weight assignment, respectively. The regression subnet is used for size and position prediction of bounding boxes, and the classification subnet is used for confidence scores prediction of bounding boxes.

Figure 2. The framework of the proposed object detection algorithm. It mainly includes the backbone, feature pyramid, adaptive weight assignment based on dual attention, regression, and classification. The adaptive weight assignment module based on dual attention involves both spatial weight and channel weight.

Regression subnet. The regression subnets are attached to every feature level to predict the corresponding bounding boxes. A regression subnet consists of four $3 \times 3$ convolutional layers and outputs the feature maps with four channels for regression. Take the $i$th level of feature pyramid as an example, we can generate the feature maps of shape \((4, H_i, W_i)\) for regression. \((H_i, W_i)\) is the size of feature maps, and four channels indicate the distances from one specific position to the four boundaries of a bounding box. Given a position and its corresponding four values, a candidate bounding box can then be obtained.

Classification subnet. The classification subnets are also attached to every feature level to estimate the classification scores. For the $i$th level of feature pyramid, we can generate the feature maps with shape \((C, H_i, W_i)\) for classification. $C$ denotes the channel number as well as the category number, which is 80 for the MS-COCO dataset [27]. On the feature maps for classification, if one position falls into a ground-truth bounding box, it can be regarded as the positive sample, otherwise it is the negative sample. Moreover, it is worth noting that at the high level of feature pyramid, some small bounding boxes may never appear in classification output. Accordingly, following FCOS [13], we assign different ground-truth bounding boxes to different levels of feature pyramid. Specifically, for the five levels \(\{P_3, P_4, P_5, P_6, P_7\}\), we define a set of threshold values \(\{m_2, m_3, m_4, m_5, m_6, m_7\}\) as \(\{0, 64, 128, 256, 512, +\infty\}\). If \(m_{i-1} < \max(l, r, t, b) < m_i\), where \(\max(l, r, t, b)\) represents the...
maximum distance to the boundaries of a certain ground-truth bounding box, the certain ground-truth bounding box will be assigned to the $i$th feature level.

### 3.2. Adaptive Weight Assignment Based on Dual Attention

The structure of adaptive weight assignment module based on improved dual attention mechanism is shown in Figure 3. It consists of two components: (a) spatial weight attention and (b) channel weight attention. The adaptive weight assignment module is added right before the classification subnet so that the confidence scores can be adjusted with the adaptive weights.

![Figure 3](image)

**Figure 3.** The structure of adaptive weight assignment module based on improved dual attention. It is mainly composed of two parts: (a) spatial weight attention and (b) channel weight attention. DW Conv means the depthwise convolution. Avg Pool and Max Pool mean average pooling and max pooling, respectively. GAP means global average pooling. $\sigma$ means the sigmoid activation.

Spatial weight attention. In Figure 3, the part of (a) represents the spatial weight attention. To obtain the spatial weights, existing algorithms usually employ the standard convolution. However, standard convolution involves the interaction in both spatial and channel dimensions, which will lead to the spatial weights being confused with the channel weights. Hence we propose to take the depthwise convolution in which each convolution kernel only operates on one channel of feature map. By separating spatial weight attention and channel weight attention thoroughly, the depthwise convolution can help learn a more reasonable distribution of confidence scores. Assume the input feature maps have the shape of $(C, H, W)$, then the output of depthwise convolution denoted as $F^1_a$ also has the shape of $(C, H, W)$. The calculation is formulated as:

$$F^1_a(c) = V(c) * X(c),$$  

where $X(c)$, and $F^1_a(c)$ indicate the $c$th channel for input and output feature maps, respectively, and $V(c)$ is a 2D spatial kernel for the $c$th channel.

After depthwise convolution, pooling operation is utilized to further extract salient features. We adopt two types of pooling operations, i.e., the average pooling and max pooling, to process the $F^1_a$, respectively. The output feature maps of average pooling and
max pooling have the same shapes, and are concatenated together to obtain the new feature maps $F_2^a$ as follows:

$$F_2^a = \text{Concat}(\text{MaxPool}(F_1^a), \text{AvgPool}(F_1^a)),$$

where $\text{MaxPool}$ and $\text{AvgPool}$ refer to the max pooling and average pooling, and $F_2^a$ has the shape of $(2C, H, W)$. Afterwards, the $F_2^a$ is transformed to $F_3^a$ with the shape of $(C, H, W)$ through a $1 \times 1$ convolutional layer.

The operation of pooling especially the max-pooling usually loses some valuable information of feature maps, e.g., the abundant spatial relationship between local regions. To compensate for the lost information, the residual structure is needed. Hence we add the $F_1^a$ and $F_3^a$ to generate the new feature maps $F_4^a$. The $F_4^a$ is lastly processed by a sigmoid function to make the value of each position on feature maps limited between 0 and 1. With the feature maps of spatial weights, we just multiply it by the $F_1^a$ to obtain the output $F_5^a$.

Channel weight attention. Since different information may exist across different channels, the interaction in channel dimension is also vital for attention estimation. As shown in Figure 3, the part of (b) indicates the channel weight attention, which aims to calculate the reasonable weights distribution across channels. Specifically, the GAP means the global average pooling. Following SENet [39] and ECANet [22], to learn the weight information about channels, the global average pooling needs to be firstly performed as follows:

$$F_1^b = g(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{ij},$$

where $H$ and $W$ represent the height and width of input feature maps $X$, and $F_1^b$ has the shape of $(C, 1, 1)$.

Next, the Conv1d which means the 1D convolutional layer is utilized to learn the relations between different channels. Recently, the group convolution has been extensively used in various classification and detection networks with great performance gain. Hence, following [22], we opt to take the 1D convolutional layer instead of fully connected layer for better performance and fewer parameters. Concretely, the $F_1^b$ is fed into a 1D convolutional layer with a fixed kernel size of 5. The output is further processed by a sigmoid activation function to constrain the values to be between 0 and 1, obtaining $F_2^b$. The formula can be written as:

$$F_2^b = \sigma(\text{Conv1d}(F_1^b)) = \sigma(\text{Conv1d}(g(X))),$$

where $\sigma$ refers to the sigmoid activation function. At last, the $F_3^a$ for spatial weight attention is multiplied by the $F_2^b$ for channel weight attention, obtaining the final feature maps with shape of $(C, H, W)$ for the adaptive weight assignment module.

Stacked dual attention. We stack the multiple modules of dual attention to further improve the performance. As shown in Figure 4a, the classification branch commonly used in existing algorithms [44] is a fully convolutional network, which consists of four standard convolutional layers with the last one for score calculation. We replace the first three standard $3 \times 3$ convolutional layers with the three stacked dual attention modules, constituting a new classification branch, as shown in Figure 4b. For fair comparison, the same layers of original classification branch and modified classification branch have the same size of output feature maps.
Feature Map (1)  |  Feature Map (1)  |  Feature Map (1)  |  Feature Map (1)  
Feature Map (2)  |  Feature Map (2)  |  Feature Map (2)  |  Feature Map (2)  
Feature Map (3)  |  Feature Map (3)  |  Feature Map (3)  |  Feature Map (3)  
Feature Map (4)  |  Feature Map (4)  |  Feature Map (4)  |  Feature Map (4)  

(a) Original Classification Branch  
(b) Classification Branch with Attention

Figure 4. The comparison of (a) original classification branch and (b) modified classification branch with stacked dual attention. In (a), the standard convolutional layers are used. In (b), three standard convolutional layers are replaced by the three stacked dual attention modules.

3.3. Loss

Our overall loss for training is

\[ L = \lambda_1 L_{\text{reg}} + \lambda_2 L_{\text{cls}}, \]  

(5)

where \( L_{\text{reg}} \) and \( L_{\text{cls}} \) denote the regression loss and classification loss separately, \( \lambda_1 \) and \( \lambda_2 \) are the corresponding weights and are both set as 1 in our experiments.

We take the IoU loss for regression. The \( L_{\text{reg}} \) can then be written as

\[ L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{H} \sum_{j=1}^{W} -\log(\text{IoU}(P_{ij}, Y_{ij})), \]

(6)

where \( N \) indicates the number of positive samples, \( P_{ij} \) and \( Y_{ij} \) are predicted bounding box and ground-truth bounding box on position \((i, j)\), respectively, and \( \text{IoU} \) means the operation of intersection over union. Note that only the positive samples are counted. The regression loss for negative samples is immediately set to 0.

We use the focal loss [30] for classification. The \( L_{\text{cls}} \) can be written as

\[ L_{\text{cls}} = \frac{1}{N} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} \left\{ -\theta(1 - P_{ijc})^\gamma \log(P_{ijc}), \quad \text{if } Y_{ijc} = 1 \\
- (1 - \theta) P_{ijc}^\gamma \log(1 - P_{ijc}), \quad \text{otherwise} \right. \]

where \( N \) is the total number of samples including positive and negative samples, \( P_{ijc} \) and \( Y_{ijc} \) are predicted score and ground-truth score, respectively, and \( \theta \) and \( \gamma \) are hyper-parameters which are set to 0.25 and 2 separately. The focal loss concentrates more on the easily misclassified samples, thus can alleviate the samples imbalance issue and gain better performance.

4. Experiment

The proposed algorithm in this paper is evaluated on the MS-COCO dataset [27] which includes over 1.5 million images with 80 categories. Section 4.1 gives training and testing details. Section 4.2 introduces visualized analysis for our proposed algorithm adopting adaptive weight assignment strategy. Section 4.3 presents the results of experiments.
4.1. Implementation Details

Training details: The backbone networks utilized in the proposed algorithm for training include ResNet-50 and ResNet-101 [42], which are pretrained on the ImageNet dataset. The maximum number of iterations is set as 200 K, and the initial learning rate is set to $5 \times 10^{-3}$. The learning rate will be divided by 10 each time up to the 120 K and 160 K iterations. As for the input image, we resize it to 800 on the shorter side and less than or equal to 1333 on the longer side. The proposed algorithm is trained on 2 TitanX GPUs with a batch size of 8. On the MS-COCO [27] dataset, the split called trainval35K is used for training, which contains 80 K images for training set and 35 K images for validation set.

Testing details: After obtaining the predicted bounding boxes with specific categories through regression and classification subnets, we utilize the technique of non-maximum suppression (NMS) [45] to filter out the redundant predicted bounding boxes with high Intersection over Union (IoU) but low confidence score. The experimental results are evaluated through six metrics, namely AP, AP$_{50}$, AP$_{75}$, AP$_S$, AP$_M$ and AP$_L$. AP is averaged over multiple IoU values and the IoU thresholds are from 0.5 to 0.95, which acts as the main evaluation indicator. AP$_{50}$ and AP$_{75}$ refer to IoU = 0.5 and 0.75, respectively. It is obvious that the larger the IoU threshold, the stricter the corresponding evaluation metric. AP$_S$, AP$_M$ and AP$_L$ represent the average precision over different sizes of objects: area < $32^2$, $32^2$ < area < $96^2$ and area > $96^2$.

4.2. Visualized Analysis

As mentioned in Section 3, the proposed improved dual attention mechanism is added before the classification subnet to obtain adaptive weight distribution. In this section, we take an example to visualize the generated confidence scores for classification with different weight distributions and further analyze the impact on object detection.

As shown in Figure 5, an athlete in input image is to be detected. Figure 5a shows the confidence scores for classification without any weight assignment. Figure 5b illustrates the weight distribution strategy named centerness [13], which assigns the center area of bounding box with high weight and the marginal area with low weight. Figure 5c shows the classification confidence scores with centerness strategy, which is obtained through multiplying the original scores by weights from centerness. Figure 5d shows the classification confidence scores with our proposed attention mechanism.

![Figure 5](image-url)

Figure 5. The visualization of classification confidence scores with different weight distributions. The left shows input image. (a) Shows the original classification confidence scores without weight assignment. (b) Shows the weight distribution of centerness. (c) Shows the classification confidence scores with weight distribution of centerness. (d) Shows the classification confidence scores with our proposed attention mechanism.
It can be seen that in Figure 5a, almost the whole region of athlete gains similar high scores, including the marginal area. With the high scores, some low-quality bounding boxes predicted at positions far away from the center of an object will also be preserved, leading to the detection performance degradation. In Figure 5c that shows the classification scores with centerness, the positions far away from central area tend to have lower confidence scores, which will restrain the low-quality bounding boxes to some extent. However, the central area is not as distinguishable as areas such as the head for athlete in input image, and does not deserve the highest score. In Figure 5d that shows the classification scores with our proposed attention mechanism, the area with highest score has an adaptive central deviation, and focuses on the athlete head that is the most recognizable part. Hence the results of Figure 5 demonstrate that our proposed dual attention mechanism does implements the adaptive weight assignment strategy, and can help improve the quality of predicted bounding boxes.

4.3. Experimental Results

In this section, the experimental results for verifying effectiveness of improved dual attention, ablation study and comparison with the state-of-the-arts are displayed and analyzed.

4.3.1. Effectiveness of Improved Dual Attention

Different from the original dual attention mechanism, our proposed improved version adopts the depthwise convolution instead of standard convolution for spatial weight attention, and the 1D convolution instead of fully connected layer for channel weight attention. The effectiveness of the two modifications are verified separately below.

Firstly, for spatial weight attention, we take ResNet-50 and ResNet-101 as the backbone network in turn, and compare the standard convolution with the depthwise convolution under serial and parallel dual attention frameworks, respectively. The results are shown in Table 1. It can be clearly seen that the parallel framework is slightly better than the serial framework, and the depthwise convolution greatly improves the standard convolution, whether using ResNet-50 or ResNet-101. In our analysis, the depthwise convolution brings more performance gains than the parallel framework, which may be because the former can distinguish the spatial and channel attentions more thoroughly, and solve the issue of mutual interference more effectively.

Table 1. Comparison between standard convolution and depthwise convolution under serial and parallel dual attention frameworks. The backbones we used are ResNet-50 and ResNet-101.

| Method                    | Framework | Backbone   | AP  |
|---------------------------|-----------|------------|-----|
| standard convolution      | Serial    | ResNet-50  | 45.0|
| depthwise convolution     | Serial    | ResNet-50  | 46.7|
| standard convolution      | Parallel  | ResNet-50  | 45.2|
| depthwise convolution     | Parallel  | ResNet-50  | 47.1|
| standard convolution      | Serial    | ResNet-101 | 49.0|
| depthwise convolution     | Serial    | ResNet-101 | 52.2|
| standard convolution      | Parallel  | ResNet-101 | 49.3|
| depthwise convolution     | Parallel  | ResNet-101 | 52.7|

Secondly, for channel weight attention, we compare the fully connected layer and the 1D convolution by using ResNet-50 as the backbone network and adopting the parallel framework. The results of AP and Frames-per-Second (FPS) are listed in Table 2. It shows that the 1D convolution achieves a significant performance improvement over the fully connected layer, reaching 4.4%. Such disparity in performance may be because the fully connected layer contains many more parameters and tends to suffer the overfitting problem.
Meanwhile, the inference speed of 1D convolution is much faster than fully connected layer. Hence the 1D convolution is adopted for both higher efficiency and better performance.

### Table 2. Comparison between fully connected layer and 1D convolution for channel weight attention.

| Method            | Backbone | AP   | FPS |
|-------------------|----------|------|-----|
| fully connected layer | ResNet-50 | 42.7 | 9.8 |
| 1D convolution     | ResNet-50 | 47.1 | 13.3|

#### 4.3.2. Ablation Study

Loss weights: Our overall loss for training is composed of regression loss and classification loss. To determine the appropriate weights, we experiment with different $\lambda_1$ and $\lambda_2$. The results of proposed algorithm taking ResNet-50 as backbone on the MS-COCO benchmark are shown in Table 3. It can be seen that the best performance is achieved by $\lambda_1 = 1.0$ and $\lambda_2 = 1.0$. Imbalanced loss weights cause the significant performance degradation. Hence we set $\lambda_1$ and $\lambda_2$ both to 1.

### Table 3. Comparison between weights for regression loss and classification loss.

| $\lambda_1$ | $\lambda_2$ | AP   |
|-------------|-------------|------|
| 1.0         | 1.0         | 47.1 |
| 1.0         | 2.0         | 43.6 |
| 2.0         | 2.0         | 43.5 |

Framework design: There are two frameworks, i.e., the serial and parallel frameworks, that can be used for the proposed improved dual attention mechanism. We adopt them, respectively, for comparison. The experimental results are shown in Table 4. It demonstrates that both spatial weight attention and channel weight attention contribute a lot to the performance improvement, whether using serial or parallel framework. In addition, the channel weight attention plays a good complementary role to the spatial weight attention and effectively improves the performance (5.6% under serial framework and 5.2% under parallel framework), which indicates that information exchange between the channels is also significant for representation learning.

### Table 4. Comparison between different frameworks for proposed improved dual attention module.

| Channel Weight | Spatial Weight | Framework | AP |
|----------------|---------------|-----------|----|
| ✓              | ✓             | Serial    | 33.1|
| ✓              |               | Serial    | 41.3|
| ✓              | ✓             | Serial    | 46.9|
| ✓              |               | Parallel  | 33.5|
| ✓              | ✓             | Parallel  | 41.9|
| ✓              | ✓             | Parallel  | 47.1|

#### 4.3.3. Comparison with the State of the Art

Our proposed algorithm is evaluated on the MS-COCO benchmark to make comparison with the state-of-the-art anchor-based and anchor-free object detection algorithms, as shown in Table 5. It can be seen that our anchor-free algorithm achieves the results of 47.1% and 52.7% with the backbone of ResNet-50 and ResNet-101, respectively. Ours with backbone of ResNet-101 even outperforms most of the anchor-based detectors, and is competitive with YOLOv4-P7 [46] (55.5%) and EfficientDet-D7 [47] (53.7%). Both the two algorithms use the particular backbones for their own network structures, while ours adopts the common backbone, i.e., ResNet, and will be probably improved as the backbone network is updated. For anchor-free object detection, ours with backbone of ResNet-101
defeats all the other algorithms, and achieves the substantial improvements. For example, ours is better than the best keypoint-based algorithm RepPoints [11] with a 7.7% increase in accuracy (52.7% vs. 45.0%). The comparison results with state-of-the-art detectors fully demonstrate the effectiveness of our proposed algorithm with adaptive weight assignment strategy and modified parallel spatial and channel weight attention framework.

Table 5. Comparison with the state-of-the-art detectors on MS-COCO benchmark. The backbones include ResNet (R) [42], DetNet (D) [48], ResNeXt (X) [49], Dual Path Network (DPN) [50], Cross Stage Partial (CSP) [51], EfficientNet (E) [52], and Hourglass (H) [53].

| Method              | Backbone     | AP  | AP50 | AP75 | APs | APm | API  |
|---------------------|--------------|-----|------|------|-----|-----|------|
| Anchor-based detectors:                |              |     |      |      |     |     |      |
| Faster R-CNN [8]    | R-101        | 36.2| 59.1 | 39.0 | 18.2| 39.0| 48.2 |
| DetNet [45]         | D-59         | 40.3| 62.1 | 43.8 | 23.6| 42.6| 50.0 |
| Soft-NMS [54]       | R-101        | 40.8| 62.4 | 44.9 | 23.0| 43.4| 53.2 |
| C-Mask R-CNN [55]   | R-101        | 42.0| 62.9 | 46.4 | 23.4| 44.7| 53.8 |
| Cascade R-CNN [29]  | R-101        | 42.8| 62.1 | 46.3 | 23.7| 45.5| 55.2 |
| Libra R-CNN [56]    | X-101-64×4d  | 43.0| 64.0 | 47.0 | 25.3| 45.6| 54.6 |
| Revisiting R-CNN [57]| R-101+R-152  | 43.1| 66.1 | 47.3 | 25.8| 45.9| 55.3 |
| SNIP [58]           | DPN-98       | 45.7| 67.3 | 51.1 | 29.3| 48.8| 57.1 |
| TridentNet [59]     | R-101-DCN    | 46.8| 67.6 | 51.5 | 28.0| 51.2| 60.5 |
| RefineDetS12 [60]   | R-101        | 36.4| 57.5 | 39.5 | 16.6| 39.9| 51.4 |
| RetinaNet [30]      | R-101        | 39.1| 59.1 | 42.3 | 21.8| 42.7| 50.2 |
| YOLOv4 [7]          | CSPDarkNet-53| 43.5| 65.7 | 47.3 | 26.7| 46.7| 53.3 |
| YOLOv4-P7 [46]      | CSP-P7       | 55.5| 73.4 | 60.8 | 38.4| 59.4| 67.7 |
| EfficientDet-D7 [47]| E-B6         | 53.7| 72.4 | 58.4 | 35.8| 57.0| 66.3 |
| Anchor-free detectors: |            |     |      |      |     |     |      |
| ExtremeNet [10]     | H-104        | 40.2| 55.5 | 43.2 | 20.4| 43.2| 53.1 |
| CornerNet [14]      | H-104        | 40.5| 56.5 | 43.1 | 19.4| 42.7| 53.9 |
| CenterNet-HG [12]   | H-104        | 42.1| 61.1 | 45.9 | 24.1| 45.5| 52.8 |
| Grid R-CNN [61]     | X-101        | 43.2| 63.0 | 46.6 | 25.1| 46.5| 55.2 |
| CornerNet-Lite [31] | H-54         | 43.2| -    | -    | 24.4| 44.6| 57.3 |
| CenterNet [15]      | H-104        | 44.9| 62.4 | 48.1 | 25.6| 47.4| 57.4 |
| RepPoints [11]      | R-101-DCN    | 45.0| 66.1 | 49.0 | 26.6| 48.6| 57.5 |
| FoveaBox [62]       | X-101        | 42.1| 61.9 | 45.2 | 24.9| 46.8| 55.6 |
| FSAF [36]           | X-101-64×4d  | 42.9| 63.8 | 46.3 | 26.6| 46.2| 52.7 |
| FCOS [13] w/improv  | X-101-64×4d  | 44.7| 64.1 | 48.4 | 27.6| 47.5| 55.6 |
| SAPD [63]           | R-101-DCN    | 46.0| 65.9 | 49.6 | 26.3| 49.2| 59.6 |
| ATSS [64]           | X-101-DCN    | 47.7| 66.5 | 51.9 | 29.7| 50.8| 59.7 |
| Ours                | R-50         | 47.1| 65.7 | 51.1 | 28.2| 50.0| 58.7 |
| Ours                | R-101        | 52.7| 71.1 | 56.9 | 35.0| 56.0| 66.1 |

5. Conclusions

In this paper, we have proposed an adaptive weight assignment strategy based on improved dual attention mechanism for anchor-free object detection. It can help learn a more reasonable distribution of weights and reduce the low-quality detections by untying the dual attention effectively and efficiently, in case that the central area includes nothing of target object or is just shielded. For implementation of the dual attention mechanism, we develop a new parallel framework with the depthwise convolution for spatial attention and the 1D convolution for channel attention. The depthwise convolution can untie the spatial and channel attention thoroughly and help avoid the confusion issue. The 1D convolution can enhance the performance as well as the computational efficiency. On the MS-COCO dataset, our proposed algorithm is shown to achieve a great increment compared with other anchor-free object detectors.

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

**Funding:** This work was funded by the National Natural Science Foundation of China under Grants No. 61976010 and 62106011 and the National Key Research and Development Program of China under Grant No. 2019YFF0301805.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset analyzed during the current study is available in the MS-COCO repository [27] with the identifier accessed on 1 August 2021 “https://cocodataset.org”, which can be used with permission of COCO Consortium.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**References**

1. Zheng, Y.; Pal, D.K.; Savvides, M. Ring loss: Convex feature normalization for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5089–5097.
2. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4282–4291.
3. Tufail, A.B.; Ullah, I.; Khan, R.; Ali, L.; Yousaf, A.; Rehman, A.U.; Alhakami, W.; Hamam, H.; Cheikhrouhou, O.; Ma, Y.K. Recognition of ziziphus lotus through aerial imaging and deep transfer learning approach. *Mob. Inf. Syst.* **2021**, 1–10. [CrossRef]
4. Ahmad, I.; Ullah, I.; Khan, W.U.; Rehman, A.U.; Adrees, M.S.; Saleem, M.Q.; Cheikhrouhou, O.; Hamam, H.; Shafiq. M. Efficient algorithms for e-healthcare to solve multiobject fuse detection problem. *J. Healthc. Eng.* **2021**, 1–16. [CrossRef]
5. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
6. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* 2018, arXiv:1804.02767.
7. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* 2020, arXiv:2004.10934.
8. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single shot multiBox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
10. Zhou, X.; Zhuo, J.; Krähenbühl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 850–859.
11. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point set representation for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019; pp. 9657–9666.
12. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* 2019, arXiv:1904.07850.
13. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 9627–9636.
14. 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; pp. 734–750.
15. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint triplets for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
16. Chen, R.; Liu, Y.; Zhang, M.; Liu, S.; Yu, B.; Tai, Y.W. Dive deeper into box for object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 412–428.
17. Samet, N.; Hicsonmez, S.; Akbas, E. Reducing label noise in anchor-free object detection. In Proceedings of the British Machine Vision Conference, Manchester, UK, 7–11 September 2020.
18. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
19. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency channel attention networks. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
20. Yuan, J.; Wei, J.; Wattanachote, K.; Zeng, K.; Luo, X.; Xu, Q.; Gong, Y. Attention-Based bi-directional refinement network for salient object detection. *Appl. Intell.* **2022**, 1–13. [CrossRef]
21. Li, Y.; Zhou, S.; Chen, H. Attention-based fusion factor in FPN for object detection. *Appl. Intell.* **2022**, 1–10. [CrossRef]
22. 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 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
23. Tang, H.; Bai, S.; Sebe, N. Dual attention GANs for semantic image synthesis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1994–2002.

24. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.

25. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5658–5667.

26. Hosang, J.H.; Benenson, R.; Schiele, B. Learning non-maximum suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4307–4315.

27. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 5–12 September 2014; pp. 740–755.

28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.

29. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.

30. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.

31. Law, H.; Teng, Y.; Russakovsky, O.; Deng, J. CornerNet-Lite: Efficient keypoint based object detection. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 9–12 September 2019.

32. Rashwan, A.; Kalra, A.; Poupart, P. Matrix Nets: A new deep architecture for object detection. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.

33. Redmon, J.; Divvala, S.K.; Girshick, R.B.; 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, 26 June–1 July 2016; pp. 779–788.

34. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. DenseBox: Unifying landmark localization with end to end object detection. arXiv 2015, arXiv:1509.04874.

35. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T.S. UnitBox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.

36. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 840–849.

37. Wu, Y.; He, K. Group normalization. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.

38. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.D.; Savarese, S. Generalized intersection over union: A Metric and a loss for bounding box regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.

39. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

40. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–29 June 2019; pp. 510–519.

41. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.; Lin, H.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. arXiv 2020, arXiv:2004.08955.

42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.

43. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. 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; pp. 2117–2125.

44. Tang, Z.; Yang, J.; Pei, Z.; Song, X. Coordinate-based anchor-free module for object detection. Appl. Intell. 2021, 51, 9066–9080. [CrossRef]

45. Neubeck, A.; Gool, L.V. Efficient non-maximum suppression. In Proceedings of the International Conference on Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2006; pp. 850–855.

46. Wang, C.; Bochkovskiy, A.; Liao, H.M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 29–25 June 2021; pp. 13029–13038.

47. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2018 2020; pp. 10781–10790.

48. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. DetNet: Design Backbone for Object Detection. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 334–350.

49. Xie, S.; Girshick, R.B.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.

50. Chen, Y.; Li, J.; Xiao, H.; Jin1, X.; Yan, S.; Feng, J. Dual path networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4470–4478.
51. Wang, C.; Liao, H.M.; Wu, Y.; Chen, P.; Hsieh, J.; Yeh, I. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 390–391.

52. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.

53. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference Computer Vision, Amsterdam, the Netherlands, 11–14 October 2016; pp. 483–499.

54. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS – Improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.

55. Chen, Z.; Huang, S.; Tao, D. Context refinement for object detection. In Proceedings of the European Conference Computer Vision, Munich, Germany, 8–14 September 2018; pp. 71–86.

56. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards balanced learning for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 821–830.

57. Cheng, B.; Wei, Y.; Shi, H.; Feris, R.S.; Xiong, J.; Huang, T.S. Revisiting RCNN: On awakening the classification power of faster RCNN. In Proceedings of the European Conference Computer Vision, Munich, Germany, 8–14 September 2018; pp. 453–468.

58. Singh, B.; Davis, L.S. An analysis of scale invariance in object detection SNIP. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3578–3587.

59. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019; pp. 6054–6063.

60. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.

61. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid R-CNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7363–7372.

62. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Shi, J. FoveaBox: Beyond anchor-based object detector. IEEE Trans. Image Process. 2020, 29, 7389–7398. [CrossRef]

63. Zhu, C.; Chen, F.; Shen, Z.; Savvides, M. Soft anchor-point object detection. In Proceedings of the European Conference Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 91–107.

64. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.