electronics

MDPI

Article

LNFCOS: Efficient Object Detection through Deep Learning Based on LNblock

Beomyeon Hwang [1], Sanghun Lee [2,*] and Hyunho Han [3]

[1] Department of Plasma Bio Display, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
[2] Ingenium College of Liberal Arts, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
[3] College of General Education, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan 44610, Korea
* Correspondence: leesh58@kw.ac.kr; Tel.: +82-2-940-5287

Abstract: In recent deep-learning-based real-time object detection methods, the trade-off between accuracy and computational cost is an important consideration. Therefore, based on the fully convolutional one-stage detector (FCOS), which is a one-stage object detection method, we propose a light next FCOS (LNFCOS) that achieves an optimal trade-off between computational cost and accuracy. In LNFCOS, the loss of low- and high-level information is minimized by combining the features of different scales through the proposed feature fusion module. Moreover, the light next block (LNblock) is proposed for efficient feature extraction. LNblock performs feature extraction with a low computational cost compared with standard convolutions, through sequential operation on a small amount of spatial and channel information. To define the optimal parameters of LNFCOS suggested through experiments and for a fair comparison, experiments and evaluations were conducted on the publicly available benchmark datasets MSCOCO and PASCAL VOC. Additionally, the average precision (AP) was used as an evaluation index for quantitative evaluation. LNFCOS achieved an optimal trade-off between computational cost and accuracy by achieving a detection accuracy of 79.3 AP and 37.2 AP on the MS COCO and PASCAL VOC datasets, respectively, with 36% lower computational cost than the FCOS.

Keywords: convolution neural networks; object detection; FCOS; attention method; LNblock; lightweight

Citation: Hwang, B.; Lee, S.; Han, H. LNFCOS: Efficient Object Detection through Deep Learning Based on LNblock. *Electronics* **2022**, *11*, 2783. https://doi.org/10.3390/electronics11172783

Academic Editor: Xue (Shelley) Lin

Received: 20 July 2022
Accepted: 2 September 2022
Published: 4 September 2022

1. Introduction

The field of computer vision (CV) includes image processing tasks such as object detection [1,2], semantic segmentation [3], and super-resolution [4]. The advent of convolutional neural networks (CNNs), which are widely used deep-learning methods, has accelerated the development of CV. Object detection is a fundamental CV task used for locating and categorizing objects.

Previously, deep-learning-based object detection have used two-stage methods that employ a network each for classification and regression to achieve a high detection accuracy. These methods divide the classification and regression process into two stages based on the region of interest wherein the object exists. A representative method is the region-based convolutional neural network (R-CNN) [5–7]. However, two-stage methods have a limitation in that real-time processing is impossible because classification and regression are performed by dividing them into two stages, resulting in high computational costs. To overcome this issue, one-stage methods that achieve an optimal trade-off between detection accuracy and computation cost by performing both processes in one network have been proposed. The single-shot multibox detector (SSD) [8] and you only look once (YOLO) [9–11] methods employ this approach. With the advent of one-stage methods, object detection has exhibited real-time detection performance. However, to improve the detection accuracy, the anchor is densely used by increasing recall. This results in a serious

class imbalance problem because the foreground and background are both classified as the background during network training. To address this problem, anchor-free methods that do not predefine an anchor have been developed. Representative examples include the fully convolutional one-stage object detection (FCOS) [12] and the YOLOX [13] methods. The center-point-based anchor-free method overcomes the class imbalance problem by predicting each pixel, similar to the semantic segmentation method.

Recently, one-stage methods have shown significantly higher efficiencies compared to two-stage methods. Therefore, to obtain an optimal trade-off between computation cost and detection accuracy, lightweight methods that can identify the optimal trade-off between computation cost and hardware resources have been studied for applying deep-learning networks with limited resources within an application program. One such method is pruning, which reduces the weight of the network structure. The MobileNet method, which is used for configuring a network with minimal computational cost, has been proposed. The MobileNet [14,15] network was constructed based on depth-wise separable convolution, which is a combination of depth-wise convolution (DWConv) [16] and point-wise convolution (PWConv). Its computation cost is approximately eight times lower than the conventional standard convolutions(StdConv) and has been proposed for use in situations wherein hardware, such as mobile devices, is limited.

In this paper, we propose an anchor-free method based on FCOS, called the light next FCOS (LNFCOS), which achieves an optimal trade-off between computational cost and accuracy. The proposed method comprises feature fusion and light next block (LNblock), wherein the feature loss is minimized by considering the information of different scales through feature fusion and the standard convolution is replaced by the proposed LNblock. The proposed method can achieve the same accuracy with less computational cost compared with conventional methods because it requires a small amount of space in the channel and also performs channel information extraction. Moreover, it achieves an accuracy similar to that of conventional methods and lowers the computational cost by 46.3GFLOPs.

The main contributions of this study are as follows:

- Feature loss was minimized by combining low- and high-level information through the proposed feature fusion module;
- An optimal trade-off was achieved between computation cost and accuracy compared to the conventional methods by replacing the StdConv with an LNblock;
- A detection head that maintains computational cost and improves detection accuracy by improving the structure of the conventional detection head is proposed.

2. Related Works

2.1. Fully Convolutional One-Stage Detector

FCOS, show in Figure 1, is an anchor-free method that approaches the conventional object detection problem from the point of view of the anchor. According to the definition of the size, number, and aspect ratio of the predefined anchor size, conventional object detection methods result in a performance deviation owing to the dense prediction for the dataset when detecting objects. Additionally, the definitions of the hyperparameters related to anchors are specialized for the dataset by reflecting a heuristic point of view, which reduces the generalization performance. If the predefined anchor has a fixed size and the size of the object in the image changes significantly, object detection is difficult. To solve these problems, a method of detecting an object through a pixel unit prediction has been studied, such as the fully convolutional network (FCN) [17], called the center-point method, which has shown good performance for various tasks such as semantic segmentation and depth estimation.

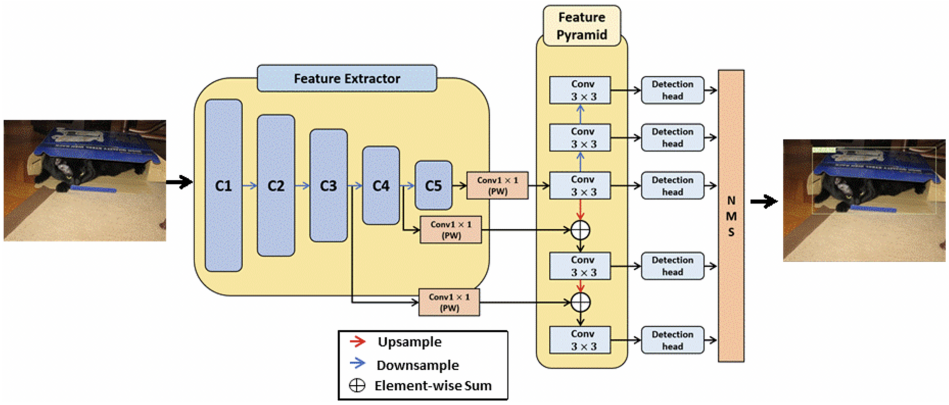

Figure 1. Structure of the fully convolutional one-stage detector. It uses ResNet-50 [18] as the backbone and has the structure of an object detection network using a feature pyramid [19] structure.

2.2. Types of Convolution Methods

With the development of deep-learning methods, convolution-related technologies have developed. Conventional convolution methods have three major limitations, which include a high computational cost, a low correlation between the channels of the feature map, and dead channels. Moreover, to accurately determine an object in an image, such as in object detection and semantic segmentation, a relatively wide receptive field must be considered to obtain sufficient contextual information. To overcome this problem, CNN methods that involve extending the receptive field by extending the kernel size of the convolution filter and using more convolutions have been used. However, these methods increase the computational cost; hence, their use is limited to some devices. Therefore, various convolution filters that are suitable for each task have been developed. For example, dilated convolution [20] extends the receptive field of the convolution kernel with a low computational cost by adding zero-padding to it; DWConv extracts spatial information from each channel at an approximately eight times lower computational cost than conventional convolutions; and deformable convolution [21] calculates a geometric pattern by learning the offset when generating an output feature map in the learning process. Figure 2 shows the structures of these convolution filters.

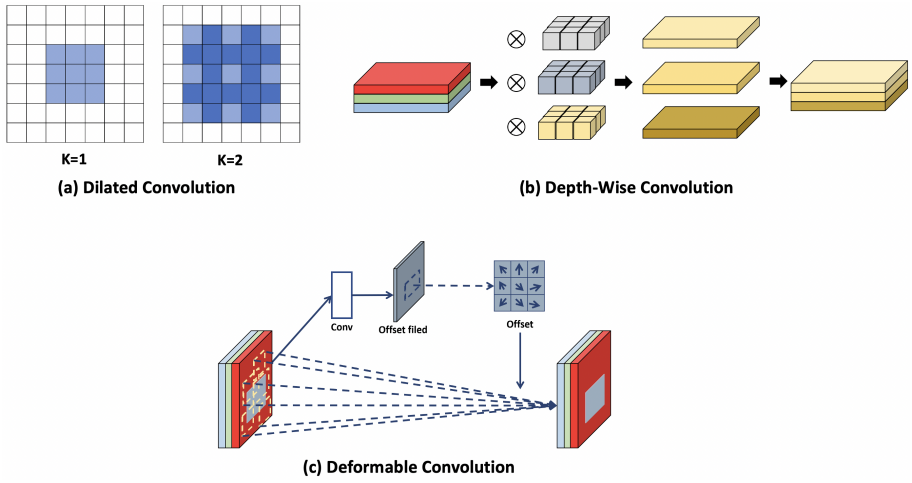

Figure 2. Structure of different convolution methods. Each convolution is a convolution that performs specialized operations to extract spatial information and geometric features, although the feature extraction performance is lower than that of the StdConv.

2.3. Attention Mechanism

The attention mechanism has been proposed to solve the long-term memory problem. It uses a fixed-length context vector of an early natural language processing model. Moreover, it emphasizes the input related to the word to be predicted at a particular point in time by re-referencing the conventional input at every point in time when predicting a word. Owing to these features, the attention mechanism has been widely used not only in the field of natural language processing, but also in the field of CV, wherein it is used to emphasize the feature information. Representative, it has been applied to each spatial channel. By calculating the correlation between the pixels within the input feature map, the importance of each pixel is identified and reflected in the output feature map by using a sigmoid activation function. With the advent of attention mechanisms, the performance of CV has improved. Representative examples of emphasis mechanisms include the convolutional block attention module (CBAM) [22] and squeeze-and-excitation module (SEM) [23]. Figure 3 shows the spatial and channel attention method of the CBAM.

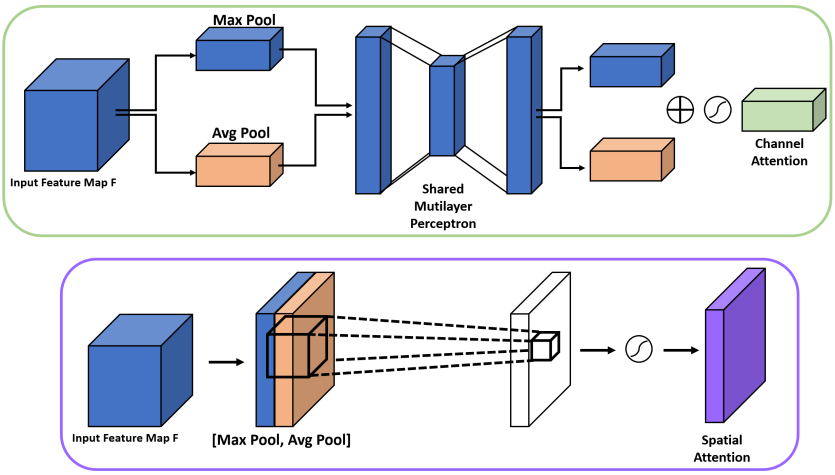

Figure 3. Channel and spatial attention method for CBAM. Each attention method used max-pooling and average-pooling to extract information from the object from the feature map and then attention to extract the important information of the channel through the sigmoid function.

3. Proposed Method

This section describes the proposed method. In Section 3.1, the structures of the proposed feature fusion module and the LNblock are described. The loss function used is described in Section 3.2.

Figure 4 shows the structure of the proposed LNFCOS. Features were extracted using ResNet-50, which is the same backbone network used in the conventional FCOS. Subsequently, to alleviate the problem of feature information imbalance, feature loss was minimized through feature fusion of low- and high-level information. In this process, the object information in the channel was emphasized by applying the channel-attention to the semantic information of each channel. Additionally, the feature pyramid was reconstructed through the LNblock, which is proposed to replace the StdConv to extract the feature information of the object with a lower computational cost. The unnecessary operation of the detection head was in addition minimized to reduce the operation cost compared to the conventional method. The details of these operations are discussed in Section 3.1.

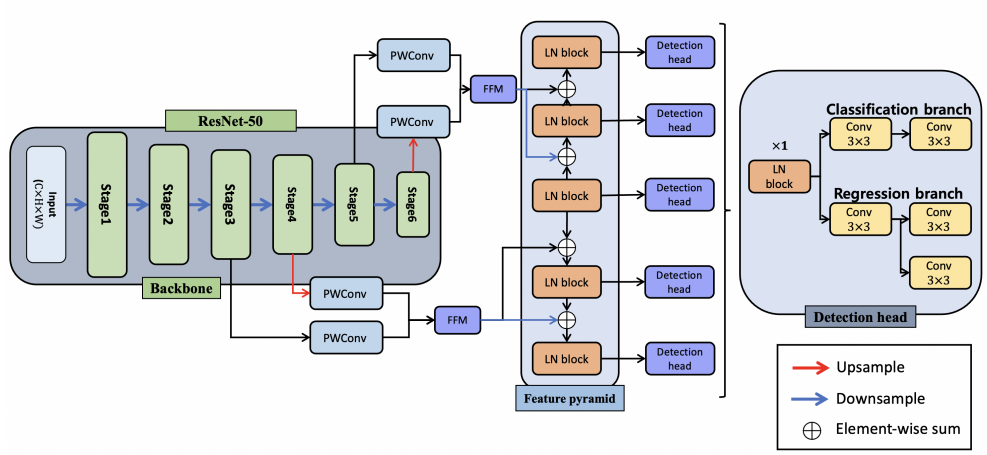

Figure 4. Structure of the proposed light next fully convolutional one-stage detector (LNFCOS). The proposed method replaces the existing 3×3 StdConv and improves spatial information by reconstructing a feature pyramid based on LNblock. In addition, based on the proposed feature fusion module, the feature information was improved with a minimum computational cost. Finally, based on the proposed LNblock, the structure of the detection head was improved to reduce the computational cost.

3.1. Network Architecture

Light next block (LNblock): The LNblock can extract spatial features with a low computational cost. It consists of a DWConv that calculates the spatial features and a PWConv that performs an operation between channels. Its structure is similar to that of the mobile bottleneck convolution (MBConv) used in MobileNet. However, the LNblock is a dilated depth-wise convolution (D2WConv), which has a wider acceptance area with the same computational cost as the conventional method and improves spatial information extraction performance with a lower computational cost compared to conventional methods. Additionally, the computational cost is minimized by minimizing the activation function and batch normalization and dividing the conventional input channel by half. Figure 5 shows the structure of the LNblock.

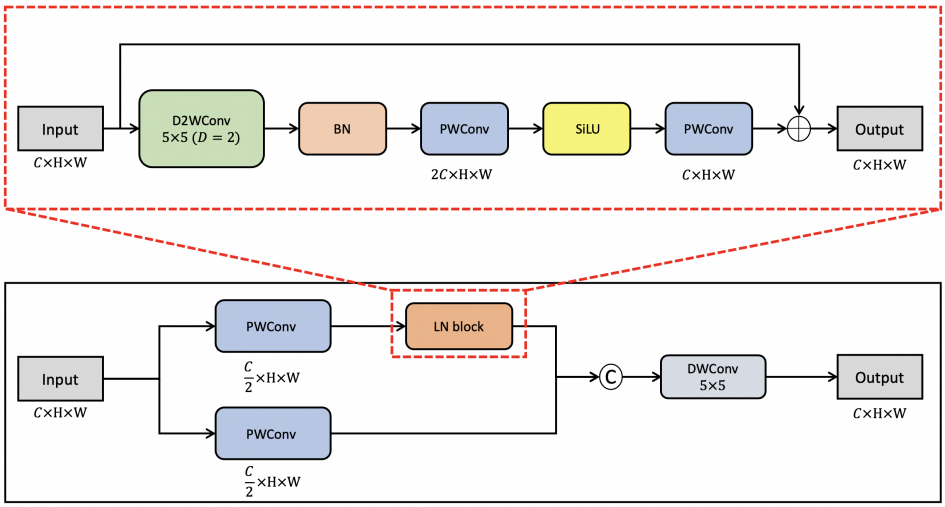

Figure 5. Structure of light next block. The proposed LNblock enables the efficient extraction of spatial information with low computational cost. Here, D2WConv, BN, PWConv, and SilU are the dilated depth-wise convolution, batch norm, 1×1 convolution, and sigmoid linear unit activation function, respectively.

First, D2Wconv, which has a smaller acceptance area than the conventional DWConv, is applied to extract spatial information. By using a wider acceptance area, it is possible to extract spatial information within a channel more efficiently compared to the conventional method with the same computational cost. Weight change is minimized by applying batch normalization to the extracted spatial information. Subsequently, the extracted spatial information is refined by expanding and contracting the channel through PWConv. Thus, the proposed LNblock can efficiently extract spatial information with a low computational cost compared to StdConv.

Feature fusion module: Figure 6 shows the proposed feature fusion module feature fusion module is a feature fusion method that minimizes loss by combining different features with a channel-highlighting technique that allows attention to contour and the semantic information required for object detection among channel information in the feature information, with a low computational cost to minimize the loss of feature information.

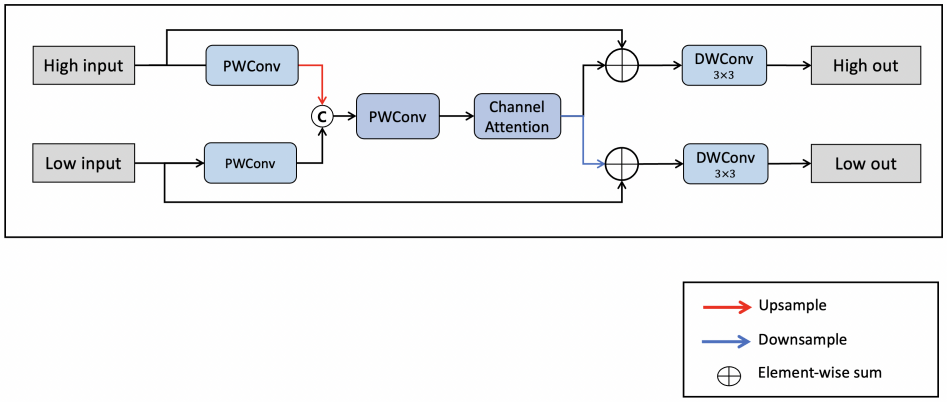

Figure 6. Structure of the feature fusion module.

The feature fusion method first uses PWConv to compress the feature information of each channel into 256 channels to integrate features of different scales. Second, it combines low- and high-level feature information by upsampling feature maps at a relatively small resolution. In addition, a channel attention method is applied to emphasize the edge and semantic information of each channel for objects in the combined feature information. Finally, the difference between conventional and emphasized information is constructed as residual learning using the input feature information. Therefore, the proposed feature fusion mitigates the imbalance between different scales. Equations (1) and (2) show the channel enhancement and the proposed feature fusion process, respectively. The symbols GAP, σ, PW_{γ},$Concat$, $F^{U}P$, and $F^{D}W$ denote global average pooling, sigmoid activation function, PWConv, concatenation operation, upsampling, and downsampling, respectively. γ denotes the expansion and contraction coefficient of the channel.

$$A_{CH}(x) = x \cdot (\sigma(PW_{\gamma}\ (PW_{\frac{1}{\gamma}}(GAP(x))))), \tag{1}$$

$$M_{FFM}(x_1,\ x_2) = A_{CH}(PW(Concat(PW(F^{up}(x_1)), PW(x_2)))). \tag{2}$$

3.2. Loss Function

The loss function of the proposed method comprises the classification and bounding box regression loss functions and the sum of the central losses. Each loss function uses focal loss [24], the generalized intersection over union (GIoU) [25], and binary cross entropy (BCE).

Classification loss: The focal loss becomes smaller than the cross-entropy loss function when P_t approaches 1. Conversely, when P_t approaches 0, the loss increases. α and γ are the hyperparameters that control the loss function, and when it is 0, it is the same

as cross-entropy. In this study, the values of hyperparameters α and γ were 2 and 0.25, respectively. Equations (3) and (4) are the cross-entropy loss and focal loss, respectively.

$$Loss_{CE}(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1-p), & \text{otherwise} \end{cases}. \tag{3}$$

where p and y represent the ground-truth (GT) and output values of the model, respectively.

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise} \end{cases},$$

$$Loss_{Focal} = -\alpha(1-p_t)^{\gamma} \cdot \log(P_t). \tag{4}$$

Regression loss: GIoU is a bounding-box regression function that performs regression based on the intersection over union (IoU), unlike the commonly used smooth L1 loss function. Equation (5) shows GIoU.

$$Loss_{GIoU}(B, B^{gt}) = 1 - IoU + \frac{\left|C - B \cup B^{gt}\right|}{|C|}. \tag{5}$$

where B, B^{gt}, and C represent the predicted bounding box, GT, and minimum size of the area covering the predicted bounding box and GT, respectively.

Finally, the centerness loss function, which gives weight to the object distance from the object center during inference, determines whether an object is conventional in the center, hence the BCE, which judges two special cases. Equation (6) represents centerness loss function.

$$Loss_{Centerness}(Y, \hat{Y}) = -(Y \log \hat{Y} + (1-Y) \log(1-\hat{Y})), \tag{6}$$

where Y and $\hat{Y}$ denote the predicted value and GT, respectively.

The final loss function of the model is shown in Equation (7) as follows:

$$Loss_{Total}(p, g) = Loss_{Focal}(p, g) + Loss_{GIoU}(p, g) + Loss_{Centerness}(p, g). \tag{7}$$

4. Experiments Results and Discussion

In this section, we show the effectiveness of the proposed method through experimental details explained in Section 4.1, a comparison of conventional networks on the PASCLVOC[26] and MS COCO [27] data sets in Section 4.2, and finally an ablation study in Section 4.3.

4.1. Implementation Details

Hyperparameter settings: In the proposed method, all data sets from PASCAL VOC and MSCOCO 2017 can be used to train and evaluate performance. Before using a dataset, we used a backbone network pretrained with the ImageNet-1K dataset.

First, the optimizer used in the datasets was stochastic gradient descent (SGD), with a momentum and weight decay of 0.9 and 0.005, respectively. For each dataset, the training batch size was 32 and 8. The number of epochs was 30. The respective input resolution of FCOS for both PASCAL VOC and MSCOCO datasets was 512×512 and 800×1333. The initial learning rate (LR) was set as 0.01. At this time, in the PASCAL VOC dataset, LR decreased by 0.1 for each iteration at 2 K and 2.1 K, and in the MS COCO, it decreased by 0.1 at 60 K and 90 K. Table 1 shows the hardware and software environment used in the experiment.

Table 1. Hardware and software environment.

Items	Descriptions
CPU	AMD Ryzen 3700X
GPU	NVIDIA RTX 3090 24 GB
RAM	64 GB
OS	Ubuntu 21.10
Framework	Pytorch 1.11

Evaluation matrix: In this study, the mean average precision (mAP), which is the average of the average precision (AP) [28] was used as an evaluation metric in object detection to evaluate the performance. The AP calculates the mAP of the recall values on the precision-recall curve (PR curve). The AP of each class was obtained with an AP value of 11 steps of the recall, 0.0, 0.1, . . . , 1.0. Equations (8)–(11) show the precision, recall, AP, and mAP.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$AP = \frac{1}{11} \sum_{r} \in 0.0, \ ..., 1.0 \rho_{interp}(r) \ , \tag{10}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \ . \tag{11}$$

where TP, FP, FN, r, ρ_{interp}, N, and AP_i denote the true positive (TP), false positive (FP), false negative (FN), recall, precision value of each recall, total number of classes, and AP value of the ith class, respectively.

4.2. Comparison with Other Networks

In this study, training and evaluation were performed on the public datasets PASCAL VOC (07+12) and MS COCO 2017.

PASCAL VOC is divided into 20 classification categories and consists of 8324 images in the training, 11,227 in the validation, and 4952 in the test datasets. In this study, we used the training and validation datasets to perform evaluation and pretraining the test datasets in the ablation studies. We compared the performances of conventional methods and the proposed method for the PASCAL VOC dataset, which is often used to compare object detection performance. The comparison was performed using an input resolution of 512×512 , which is often used in the one-stage method for PASCAL VOC.

MS COCO is divided into 80 classification categories and consists of 118,287 images in the training, 5000 in the validation, and 4952 in the evaluation datasets. In this study, it was used to compare a conventional network with other networks. The performances for the MS COCO dataset were evaluated with average values between 50% and 95% based on the threshold value of the IoU.

We also performed ablation studies to verify the performance of the proposed method and its modules.

4.2.1. PASCALVOC 2007

Table 2 shows the detection accuracy, parameters, and computation costs of the proposed and other networks for the PASCALVOC 2007 test dataset. The detection accuracy of the proposed method was higher by +0.8 AP with a lower computation cost, and it required fewer parameters than the conventional methods. Thus, the usefulness of the proposed method was confirmed.

Table 2. Comparison of other networks with PASCALVOC 07 test dataset.

Networks	Backbone	Input resolution	Parameters (M)	FLOPs (G)	mAP (%)
Two-stage					
Fast R-CNN [5]	VGG-16	600 × 1000	-	-	70.0
Faster R-CNN [6]	VGG-16	600 × 1000	134.7	-	73.2
OHEM [29]	VGG-16	600 × 1000	-	-	74.6
R-FCN [30]	ResNet-101	600 × 1000	50.9	-	80.5
One-stage					
SSD300 [8]	VGG-16	300 × 300	26.3	-	74.1
SSD512 [8]	VGG-16	512 × 512	29.4	-	76.0
YOLOv2 [10]	DarkNet-19	544 × 544	51.0	-	78.6
FCOS [12]	ResNet-50	512 × 512	32.1	103.1	78.4
LNFCOS (ours)	ResNet-50	512 × 512	27.1	60.2	79.3

Figure 7 shows the detection results of the conventional FCOS and proposed methods for the PASCAL VOC dataset. As evident from Figure 7a, through the conventional FCOS method, it is difficult to detect overlapping objects within different objects owing to the lack of spatial information when extracting features. Moreover, some objects are omitted. The results of the proposed method, shown in Figure 7b, confirm that the proposed method improves the detection accuracy with a low computational cost by improving the spatial information for images containing different objects. However, it performed some erroneous detections because of the excessive improvement in spatial information for objects with a long aspect ratio, such as the train class.

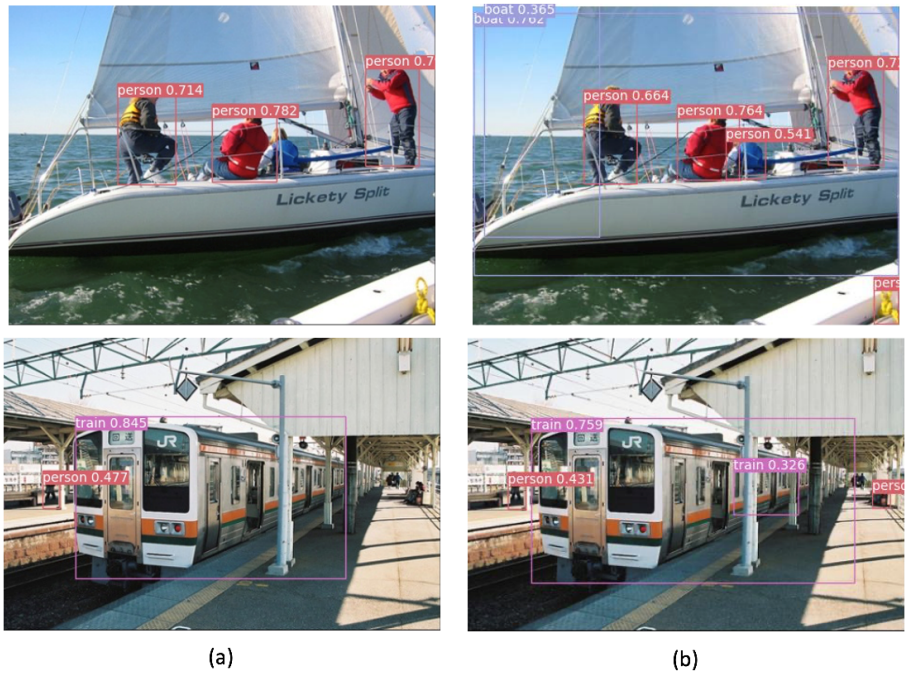

Figure 7. Detection results of (**a**) the conventional FCOS and (**b**) proposed method for the PASCAL VOC 2007 test dataset.

4.2.2. MS COCO 2017

Table 3 lists the performances of the conventional and proposed methods for the MSCOCO 2017 minival dataset. LNFCOS proceeded with a simultaneous input of 800 × 1333 as with FCOS.

The proposed method achieved a detection accuracy of approximately 37.2 AP by using ResNet-50 as the feature extraction network for the MS COCO dataset. Moreover, compared to the two-stage methods, the proposed method achieved an optimal balance

between detection accuracy and accuracy, as the number of parameters was approximately five times lower than that of Faster R-CNN. However, compared with FSAF, the AP50 decreased by −1.2 AP. The difficult object detection performance (AP75) was improved through the feature information and proposed module but decreased for AP50.

Table 3. Comparison of other networks using the MSCOCO 2017 minival dataset.

Networks	Backbone	Input Resolution	Parameters (M)	AP (%)	AP50 (%)	AP75 (%)
Two-stage						
CoupleNet [31]	ResNet-101	800 × 1024	-	34.4	54.8	37.2
FasterR-CNN [32]	ResNet-50	800 × 1024	39.8	36.7	57.3	39.3
MaskR-CNN + GRoIE [33]	ResNet-50	800 × 1333	-	38.4	59.9	41.7
One-stage						
YOLOv3 [11]	DarkNet-53	608 × 608	65.2	33.0	57.9	34.4
RetiaNet + Foveabox [34]	ResNet-50	800 × 1333	-	36.4	56.2	38.7
FSAF [35]	ResNet-50	800 × 1333	-	37.2	57.2	39.4
FCOS [12]	ResNet-50	800 × 1333	32.1	37.4	56.1	40.3
LNFCOS (ours)	ResNet-50	800 × 1333	27.1	37.2	56.0	39.9

Figure 8 shows the detection results of the conventional FCOS and the proposed method for the publicly available MS COCO dataset. The upper image in Figure 8a shows the false detection of the person, dining table, and chair class objects owing to the loss of low- and high-level information. Additionally, in the lower image, various objects are placed close to each other, and some are falsely detected owing to a lack of channel information. As shown in Figure 8b, the proposed method reduces false detection by improving feature information through the proposed feature fusion module. Additionally, the detection error is minimized by improving the characteristics of the objects by improving the spatial information, as shown in the lower image.

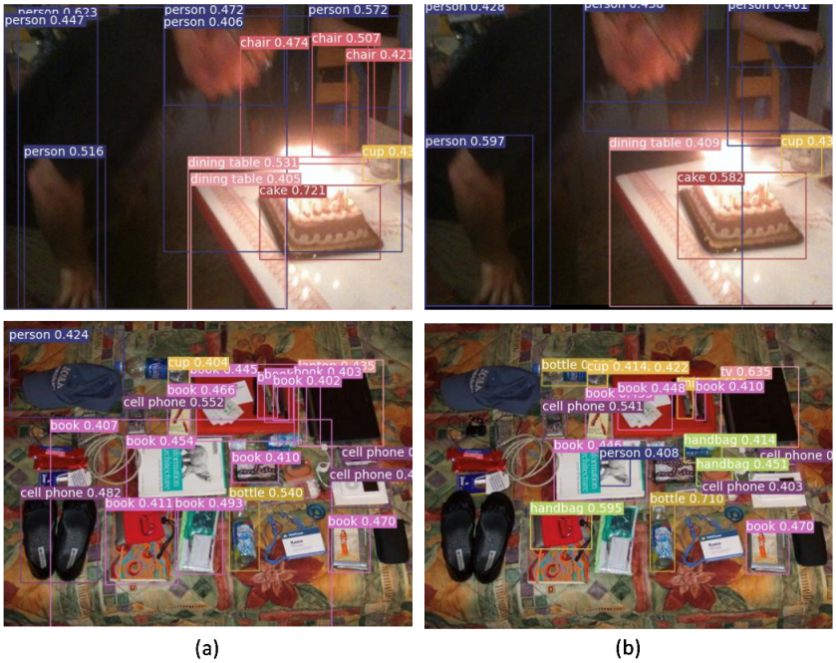

Figure 8. Detection results of (**a**) the conventional FCOS and (**b**) proposed method for the MS COCO 2017 minival dataset.

4.3. Ablation Study

To verify the effectiveness of LNFCOS, an excision study was performed using the PASCAL VOC 2007 test dataset. Table 4 lists the detection accuracy of each class of LNFCOS and FCOS for the PASCAL VOC 07 test dataset, wherein it is evident that the detection accuracy of LNFCOS is better than that of FCOS. Through the proposed LNblock, the spatial information extraction performance was improved by reconstructing the feature pyramid and detection head, and the loss of contextual information was minimized through the proposed feature fusion module to achieve an optimal balance between detection accuracy and computational cost. Compared to FCOS, LNFCOS slightly improved the detection accuracy for difficult-to-detect objects, such as boats, potted plants, and trains. However, in the case of objects with a long aspect ratio, LNFCOS showed more false detections compared with FCOS. This confirmed that there was some spatial information loss owing to the zero-padding in D2Wconv, which required a wide acceptance area and low computational cost. Through resection studies, it was confirmed that LNFCOS achieved an optimal balance between computational cost and accuracy compared with FCOS.

Table 4. The ablation study for LNFCOS analysis on the PASCAL VOC 2007 test dataset.

FCOS	**aero**	**bike**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**
	80.8	86.8	81.5	72.2	63.4	84.8	88.1	91.1	58.8	80.3
	table	**dog**	**horse**	**mbike**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**
	66.0	88.6	86.1	84.8	84.2	51.0	80.8	72.1	88.9	79.3
LNFCOS	**aero**	**bike**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**
	79.8	85.6	81.7	75.8	63.0	85.2	87.3	91.8	58.7	84.7
	table	**dog**	**horse**	**mbike**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**
	65.9	90.8	86.9	83.9	83.4	52.3	83.7	70.5	90.7	79.2

4.3.1. LNblock Analysis

In the ablation study, the parameters of the LNblock and comparisons with other methods were performed using the PASCALVOC 2007 test dataset to verify the usefulness of the LNblock.

Table 5 presents the accuracy and calculation costs of the proposed LNblock based on to the ratios of the parameters. The parameters of the LNblock comprised the kernel size K, extension factor D, and channel extension ratio C. Through the experiment, it was confirmed that the LNblock contributed to the improvement of the detection accuracy with a low computational cost. Additionally, the optimal ratio between the computational cost and detection accuracy was confirmed at a parameter ratio of (K = 5, D = 2, C = 2). Through this experiment, it was confirmed that the object detection accuracy decreased when the kernel size and expansion ratio increased more than a certain level. In the case of the channel coefficient, it was also confirmed that the detection accuracy decreased more compared with the increase in the calculation cost when the expansion ratio was large.

Table 6 presents the results of the comparative analysis between the LNblock and MBConv. MBConv has been used for efficient feature extraction at a low computational cost in EfficientNet [36], starting with MobileNet. Compared to the weight reduction applied in other methods, the computational cost of the LNblock was higher by approximately 1.9 GFLOPs compared to MBConv, but its detection accuracy was higher by 0.3 AP. However, the optimal balance between computational cost and detection accuracy was achieved by improving the detection accuracy with a lower computational cost of approximately 36.9 GFLOPs compared with the conventional StdConv.

Table 5. Ablation study results of LNblock parameter analysis for the PASCAL VOC 2007 test dataset.

Method	Parameters (M)	FLOPs (G)	mAP (%)
Baseline	32.1	103.1	78.5
LNblock (K = 3, D = 2, C = 2)	26.9	60.2	78.8
LNblock (K = 3, D = 2, C = 4)	27.1	60.8	78.6
LNblock (K = 5, D = 2, C = 2)	27.0	60.2	79.1
LNblock (K = 5, D = 2, C = 4)	27.1	60.9	78.8
LNblock (K = 7, D = 2, C = 2)	27.0	60.2	78.5
LNblock (K = 7, D = 2, C = 4)	27.2	61.0	78.3

Table 6. Ablation study results of StdConv, MBConv, and LNblock for the PASCAL VOC 2007 test dataset.

Method	Parameters (M)	FLOPs (G)	mAP (%)
Baseline	32.1	103.1	78.5
MBConv	27.1	58.1	78.8
LNblock	27.0	60.2	79.1

4.3.2. Feature Fusion Module Analysis

An ablation study was performed to confirm the accuracy and computational cost of the proposed feature fusion module. Table 7 presents the results of the proposed method with and without the feature fusion module. In the proposed method, when only the feature fusion module was used, the calculation cost increased slightly. However, it was confirmed that it improved the detection accuracy by +0.6 AP with only a slight increase in the computational cost by emphasizing the necessary feature information in each channel for low- and high-level information.

Table 7. Ablation study results of the proposed method with and without the proposed feature fusion module (FFM) for the PASCAL VOC 2007 test dataset.

Method	Parameters (M)	FLOPs (G)	mAP (%)
W/O FFM	32.1	103.1	78.8
W FFM	32.6	105.0	79.4

5. Conclusions

In this study, we proposed an efficient object detection network based on the proposed LNblock, called LNFCOS, which can achieve an optimal trade-off between accuracy and computation cost. First, the proposed method minimized the loss when extracting spatial information at a low computational cost through the LNblock. Additionally, the feature pyramid and detection head were reconstructed by replacing the standard convolution with the proposed LNblock. As a result, the features could be efficiently extracted at a lower computational cost compared to that of the conventional methods. Second, by using the proposed feature fusion module, the detection accuracy was improved by minimizing the feature loss that occurred in each channel in the conventional method, by emphasizing the feature information in the channel. For a quantitative evaluation of the proposed method, experiments and evaluations were performed on the publicly available datasets, PASCALVOC and MS COCO, to achieve an optimal balance of detection accuracy and computational cost, with a computational cost of 36% lower (43.0 GFLOPs) than that of the conventional FCOS. In addition, we confirmed the performance of the proposed method

through ablation studies. In a future study, we plan to apply the proposed method to other object detection networks and further reduce the computational cost and improve detection accuracy.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available online: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html, http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html (accessed on 20 July 2020) [26], and https://cocodataset.org/#home (accessed on 20 July 2022) [27].

Acknowledgments: The present research has been funded by the Research Grant of Kwangwoon University in 2022.

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

Abbreviations

The following abbreviations are used in this manuscript:

LNFCOS	Light next fully convolutional one-stage detector
FCOS	Fully convolutional one-stage detector
FCN	Fully convolutional network
Bn	Batch normalization
GAP	Global average pooling
IoU	Intersection of union
SiLU	Sigmoid linear unit
AP	Average precision
mAP	Mean average precision

References

1. Shin, S.; Han, H.; Lee, S. Improved YOLOv3 with duplex FPN for object detection based on deep learning. *Int. J. Electr. Eng. Educ.* **2021**, https://doi.org/10.1177/0020720920983524.
2. Park, C.; Lee, S.; Han, H. Efficient Shot Detector: Lightweight Network Based on Deep Learning Using Feature Pyramid. *Appl. Sci.* **2021**, *11*, 8692. https://doi.org/10.3390/app11188692.
3. Shin, S.; Lee, S.; Han, H. EAR-Net: Efficient Atrous Residual Network for Semantic Segmentation of Street Scenes Based on Deep Learning. *Appl. Sci.* **2021**, *11*, 9119. https://doi.org/10.3390/app11199119.
4. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
5. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
6. 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. https://doi.org/10.1109/TPAMI.2016.2577031.
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322.
8. 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 European Conference on Computer Vision. Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin, Germany, 2016; pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
9. 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, 26 June–1 July 2016; pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.

10. 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. https://doi.org/10.1109/CVPR.2017.690.
11. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
12. 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, Korea, 27 October–2 November 2019; pp. 9626–9635. https://doi.org/10.1109/ICCV.2019.00972.
13. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding Yolo Series in 2021. 2021. Available online: (accessed on 20 July 2022).
14. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
15. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140.
16. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
18. 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; IEEE Computer Society: Los Alamitos, CA, USA, 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
19. 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. https://doi.org/10.1109/CVPR.2017.106.
20. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. 2015. Available online:(accessed on 20 July 2022).
21. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 9308–9316.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
24. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324.
25. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. https://doi.org/10.1109/CVPR.2019.00075.
26. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. *The Pascal Visual Object Classes (voc) Challenge*; Springer: Berlin, Germany, 2010; Volume 88, pp. 303–338. https://doi.org/10.1007/s11263-009-0275-4.
27. 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 European conference on computer vision. Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
28. Henderson, P.; Ferrari, V. End-to-end training of object class detectors for mean average precision. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Springer: Berlin, Germany, pp. 198–213. https://doi.org/10.1007/978-3-319-54193-8_13.
29. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769. https://doi.org/10.1109/CVPR.2016.89.
30. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *2*.
31. Zhu, Y.; Zhao, C.; Wang, J.; Zhao, X.; Wu, Y.; Lu, H. CoupleNet: Coupling Global Structure with Local Parts for Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4146–4154. https://doi.org/10.1109/ICCV.2017.444.
32. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.

33. Rossi, L.; Karimi, A.; Prati, A. A Novel Region of Interest Extraction Layer for Instance Segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2203–2209. https://doi.org/10.1109/ICPR48806.2021.9412258.
34. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. FoveaBox: Beyound Anchor-Based Object Detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. https://doi.org/10.1109/TIP.2020.3002345.
35. Zhu, C.; He, Y.; Savvides, M. Feature Selective Anchor-Free Module for Single-Shot Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 840–849. https://doi.org/10.1109/CVPR.2019.00093.
36. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.