mathematics

MDPI

Article

Multi-Scale Feature Selective Matching Network for Object Detection

Yuanhua Pei [1], Yongsheng Dong [1,*], Lintao Zheng [1], Jinwen Ma [2]

[1] School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China; peiyuanhua2020@163.com (Y.P.); zhenglintao@126.com (L.Z.)
[2] Department of Information and Computational Sciences, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China; jwma@math.pku.edu.cn
* Correspondence: dongyongsheng98@163.com

Abstract: Numerous deep learning-based object detection methods have achieved excellent performance. However, the performance on small-size object detection and positive and negative sample imbalance problems is not satisfactory. We propose a multi-scale feature selective matching network (MFSMNet) to improve the performance of small-size object detection and alleviate the positive and negative sample imbalance problems. First, we construct a multi-scale semantic enhancement module (MSEM) to compensate for the information loss of small-sized targets during down-sampling by obtaining richer semantic information from features at multiple scales. Then, we design the anchor selective matching (ASM) strategy to alleviate the training dominated by negative samples caused by the imbalance of positive and negative samples, which converts the offset values of the localization branch output in the detection head into localization scores and reduces negative samples by discarding low-quality anchors. Finally, a series of quantitative and qualitative experiments on the Microsoft COCO 2017 and PASCAL VOC 2007 + 2012 datasets show that our method is competitive compared to nine other representative methods. MFSMNet runs on a GeForce RTX 3090.

Keywords: deep learning; object detection; selective matchting; positive and negative sample imbalance

MSC: 68T07

Citation: Pei, Y.; Dong, Y.; Zheng, L.; Ma, J. Multi-Scale Feature Selective Matching Network for Object Detection. *Mathematics* **2023**, *11*, 2655. https://doi.org/10.3390/math11122655

Academic Editors: Jonathan Blackledge and Catalin Stoean

Received: 18 April 2023
Revised: 16 May 2023
Accepted: 29 May 2023
Published: 10 June 2023

1. Introduction

Object detection is an important task in the field of computer vision [1,2], which aims to search for the location of the object of interest in an image or video using algorithms [3]. The object detection task can be decomposed into two subtasks: localization and classification. The randomness of the spatial location of the target in the image makes it difficult for the algorithm to locate it precisely, and the variability in size and shape affects the accuracy of the category determination [4]. These reasons make it impossible to achieve object detection using a fixed scale and pose a challenge to the object detection.

In anchor-based object detection methods, the one-stage object detection methods do not need to use the selective search method to extract the region of interest (RoI), which can be predicted directly, so the one-stage methods has a better speed advantage. However, because the one-stage methods abandon selective search, a large number of negative samples are generated in the one-stage methods. After SSD [5] was proposed as a one-stage method, it received attention from many researchers. A series of YOLO methods [6–9] were proposed to bring the one-stage methods to a climax.

Although numerous one-stage methods have made great progress, they still have shortcomings in small-size object detection and positive and negative sample imbalance. To alleviate these problems, in this paper, we proposed a multi-scale feature selective matching network (MFSMNet) for object detection. MFSMNet is based on one-stage object detection methods, and the main contribution of this paper are as follows:

- In order to improve the detection performance of small-size objects, we propose a multi-scale semantic enhancement module (MSEM) architecture. The MSEM can accomplish semantic enhancement at multi-scales and enrich the semantic features of MFSMNet. The MSEM can improve the detection performance of small-size objects.
- In order to alleviate the performance constraints caused by positive and negative sample imbalance, we propose a anchor selective matching (ASM) strategy. It uses an anchor scoring mechanism to discard low-quality localized anchors as a way to alleviate positive and negative sample imbalance and improve MFSMNet detection performance.
- Our proposed multi-scale feature selective matching network (MFSMNet) has good experimental results on PASCAL VOC 2007 + 2012 and Microsoft COCO datasets, which effectively improves the performance of object detection.

The remainder of the paper is structured as follows. Section 2 describes the related work. Sectoin 3 presents our proposed approach, and the proposed network's training information is also introduced. Section 4 gives the experimental results. A brief conclusion is given in Section 5.

2. Related Work

In recent years, the development of hardware technology, especially the improvement of the performance of graphics processing units (GPUs), has provided for the possibility of fast training of deep learning models [10]. GPUs with compute unified device architecture (CUDA) provide a powerful tool for massively parallel computing power for tens of millions of parameters of deep learning models and thus for training deep learning models [11]. Therefore, deep learning-based object detection methods have become a key research direction at present. Depending on whether they use an anchor, the present deep learning-based object detection techniques can be divided into anchor-based methods and anchor-free methods.

2.1. Anchor-Free Methods

2.1.1. Keypoint-Based Methods

Keypoint-based detectors uses heatmaps to predict key points. To get the bounding box, they are divided into several groups. CornerNet [12] uses the top-left and bottom-right corner points of an object.Then the detector embeds them and predicts Predict categories and location. ExtremeNet [13] uses five key points to complete the prediction of the bounding box. Deep extreme cut [14] concatenates the heatmap with the original RGB image to form a 4-channel CNN. RepPoints [15] is the same as deformable convolutional networks [16], but the difference is that RepPoints generates a pseudo box to compare with GT to calculate the loss to regress the location of the points. CenterNet [17] adds a centroid detection branch to CornerNet and greatly improves the performance by centroid verification. CentripetalNet [18] proposes a centripetal displacement module to group corner points after corner point prediction, which reduces the false detection rate while ensuring the recall rate.

2.1.2. Center-Based Methods

In center-based methods, YOLOv1 [19] divides the image into cells and directly predicts the targets whose object centroids fall within the cells. DenseBox [20] introduces fully convolutional networks (FCN) to the field of target detection, thus enabling end-to-end detection, which directly regresses the confidence level and relative position of target occurrence. Unitbox [21] uses intersection over union (IoU) loss, and the methods is more robust to changes in scale. Due to the relatively small number of positive samples, the recall of these detectors is low. To increase the recall rate, FCOS [22], inspired by the idea of segmentation, performs object detection tasks with the help of segmentation ideas. MSRNet uses a residual structure to alleviate overfitting while improving the detection of small-sized targets [23].

2.2. Anchor-Based Methods

2.2.1. Two-Stage Methods

Two-stage methods has more stages of regions of interest (RoI) extraction than one-stage methods. R-CNN is the first algorithmic model to introduce CNN into the field of object detection [24]. Fast R-CNN [25] transforms model training from multi-stage to single-stage training by using multi-task loss. Faster R-CNN [26] uses a region proposal network (RPN) to generate the RoI. Oriented R-CNN [27] uses a full convolutional structure to reduce the number of parameters. By Mask R-CNN [28] found pixel bias in the RoI Pooling layer, so it used bilinear interpolation to replace the RoI Pooling layer with the RoI Align layer to achieve better detection. In addition, the mask head uses a top-down approach for segmentation [29–31]. The use of region proposals with appropriate ORPs for training increased the detection performance of irregular objects in [32]. Additionally, the method suppresses error detection using a model-driven algorithm.

2.2.2. One-Stage Methods

The one-stage methods directly classify and regress the pre-defined anchor to accomplish the target detection task [33]. SSD [5] uses feature maps from several different convolutional layers to classify and regress the anchor with different step sizes. DSSD [34] uses the residual module to further extract the depth features for regression and classification based on SSD. YOLOv2 [6] adds a batch normalization (BN) layer after each convolutional layer to solve the problem of gradient disappearance and explosion while reducing hyperparameters, and YOLOv2 [6] also further takes into account fine-grained features. DarkNet-53 proposed by YOLOv3 [7] further reduces the number of network layers with similar accuracy as ResNet-101. RetinaNet [35] proposes focal loss to improve detection accuracy. RefineDet [36] filters and eliminates negative samples by introducing the anchor refinement module, thus alleviating the positive and negative sample balance. In [37], the method focuses on objects with high ORP and provides a straightforward method to enhance their detection.

3. Our Proposed Method

3.1. Network Structure Design

We propose multi-scale feature selective matching network (MFSMNet) for object detection using a residual network (ResNet) as the backbone. The features are extracted from the last three layers of the backbone. Then, we use the features in the multi-scale semantic enhancement module (MSEM) structure. The semantic enhancement module (SEM) at different scales can be used to further enrich the multi-scale and multi-spatial semantic information, and the feature maps processed by the MSEM structure are fed to the detection head. The detection head completes the classification and regression tasks according to the anchor selective matching (ASM) strategy to achieve object detection. The MFSMNet is shown in Figure 1.

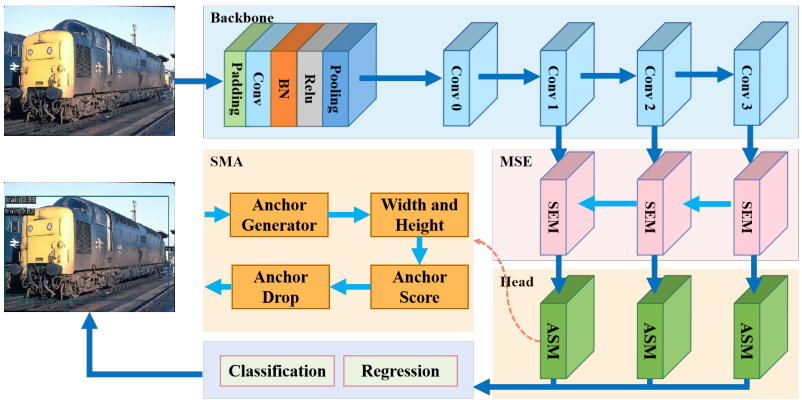


Figure 1. Architecture diagram of multi-scale feature selective matching network.

The mathematical expression of backbone network output features of multi-scale feature selective matching network can be defined as:

$$P_k = f_{resnet}(P_{in}, k), k \in \{n-2, n-1, n\} \tag{1}$$

where P_k is the backbone ResNet-50 output feature, f_{resnet} for the feature extraction operation of ResNet, P_{in} is the input image, k is the number of backbone output feature layers, and n is the maximum number of network layers of the backbone ResNet. The output feature of the anchor selective matching (ASM) strategy can be defined as:

$$P_i = f_{asm}(f_{msem}(P_k, i), i), i \in [1, 3] \tag{2}$$

where P_i is the output feature of ASM, f_{asm} is the anchor matching performed by ASM strategy, f_{msem} is the semantic enhancement operation of the MSEM structure, and i is the number of output feature layers of ASM. The output image can be defined as:

$$P_{out} = \left[f_{cls}\left(smooth \sum_{i=1}^{3} P_i \right), f_{reg}\left(smooth \sum_{i=1}^{3} P_i \right) \right] \tag{3}$$

where P_{out} is the output image of MFSMNet, f_{cls} is the classification operation branch, f_{reg} is the localization operation branch, *smooth* is the feature smoothing operations, and P_i is the output feature of ASM.

3.2. Multi-Scale Semantic Enhancement Module

Because the use of down-sampling structure in a convolutional neural network (CNN) provides richer semantic information, medium-sized and large-sized targets have more pixel points than small-sized objects. During down-sampling, the number of convolutions for small-sized targets is much smaller than for other-sized targets. This results in less valuable information about small-sized objects in the deep feature maps. In addition, as small objects carry limited information on their own, this makes it difficult for the feature extraction stage to obtain more effective features on small-sized targets. Therefore, we propose a multi-scale semantic enhancement module (MSEM).

The multi-scale semantic enhancement module (MSEM) architecture consists of five scales of semantic enhancement module (SEM). The features extracted from the backbone are fed into the semantic enhancement module, and the input features are fed into the detection head for the prediction task by two branches of spatial pyramid pooling (SPP) and depthwise separable convolution (DSC) after concat and other operations. In addition, to further mitigate gradient disappearance or gradient explosion, a residual structure is introduced. The structure of SEM is shown in Figure 2.

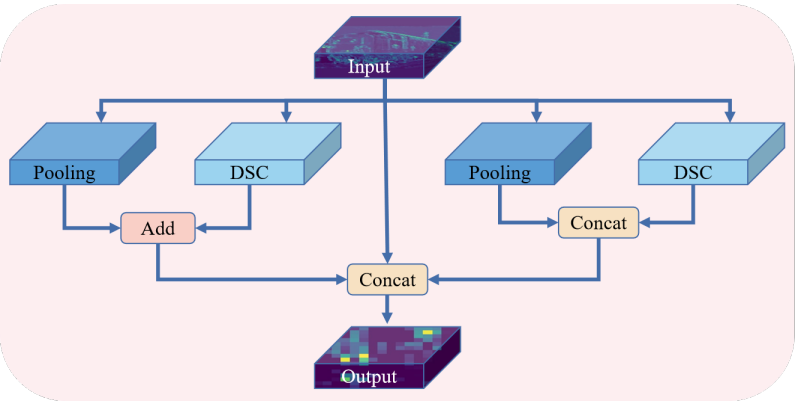


Figure 2. Architecture diagram of SEM.

The mathematical expression of the multi-scale semantic enhancement structure can be defined as:

$$F_{i+1} = Concat(Add(DSC(F_i), P(F_i)), Concat(Concat(DSC(F_i), P(F_i))) \quad (4)$$

where F_i is the input feature, F_{i+1} is the output feature, DSC denotes depth-separable convolution, P denotes pooling operation, Add denotes feature fusion, and $Concat$ denotes channel connection.

The effective information of small-size targets is lost as the network layers are deepened during down-sampling. Therefore, nearest neighbor up-sampling is used in multi-scale semantic enhancement module (MSEM) to make the size of the feature map rich in semantic information larger, and the larger size feature map is used to detect small-size objects. At the same time, we introduced spatial pyramid pooling (SPP) and depthwise separable convolution (DSC) in the semantic enhancement module (SEM). Spatial pyramid pooling allows pooling of image blocks of different sizes into fixed size feature vectors. It obtains multiple feature vectors at different scales by constructing image pyramids at different scales, down-sampling the images multiple times, and then performing pooling operations on the images at each scale. Deep separable convolution is a lightweight convolutional neural network (CNN) structure, which divides the standard convolution operation into two steps: deep convolution and point-by-point convolution. DSC can greatly reduce the number of parameters and computational effort while ensuring high accuracy. In order to enhance the semantic information as much as possible, two feature enhancement operations, concat and add, are used in the SEM. Thus, the MSEM structure can be applied to five-scale feature maps to improve accuracy while minimizing computational effort, thus achieving semantic enhancement of multi-scale features.

3.3. Anchor Selective Matching Strategy

Since the feature pyramid network [38] (FPN) methods have been proposed, many methods often use FPN structures to implement multi-scale object detection tasks due to the multi-scale feature maps of FPN. In addition, after combining the manual design of an anchor, usually the lower layer of a large-size feature map in FPN is matched with a small-size anchor to match the small-size target, and the upper layer of a small-size feature map is matched with a large-size anchor. This is because the small-size feature map in the upper layer of FPN has more semantic information and is suitable for the detection of large-size objects, whereas the large-size feature map in the lower layer has more fine-grained information and is suitable for the detection of small-size objects. This design creates two limitations that limit the detection performance of the object detection task: One is that the matching mechanism of the anchor is a heuristic, which leads to non-optimal matching of each feature at training time. Second, the judging metrics used in previous non-maximum suppression (NMS) are intersection over union (IoU), which is simple and intuitive, but it only considers the overlapping area of two frames. Therefore, we propose the anchor selective matching (ASM) strategy to alleviate the performance constraints caused by the use of fixed threshold matching in existing methods.

First, the scoring mechanism of the anchor quality is defined, and the quality of the anchor is reflected by the classification effect and regression effect. Therefore, the scoring of anchor quality is shown below:

$$S_{anchor} = S_{cls} \times S_{reg} \quad (5)$$

where S_{anchor} is the anchor score, S_{cls} is the anchor category score (category probability), and S_{reg} is the anchor positioning score. However, in the output header of the network, the regression task outputs the value of the encoded offset, and not the regression score. Therefore, we introduce $DIoU$ into the output head of the network by the offset output

from the regression task. The results of $DIoU$ were used as regression quality scores as shown below:

$$DIoU = IoU - \frac{\rho^2(b, b_{gt})}{c^2} \tag{6}$$

where $DIoU$ is the conversion of ordered pairs of regression offset values into constants, b is the center point of the prediction bounding box, b_{gt} is the center point of the true bounding box, and c is the diagonal length of the smallest enclosing box covering the two boxes. However, the range of DIoU is $[-1, 1]$. Therefore, S_{reg} is defined as:

$$S_{reg} = \begin{cases} DIoU, DIoU > 0 \\ 0, DIoU \leq 0 \end{cases} \tag{7}$$

The regression branch scoring mechanism is shown in Equation (7). The anchor selective matching (ASM) strategy discards low quality anchors. Then, the number of negative samples will be massively reduced. Therefore, ASM can improve the detection performance of the network by avoiding the training dominated by negative samples as much as possible.

4. Experimental Result

4.1. Datasets and Metrics

The PASCAL VOC [39] and Microsoft COCO [40] datasets are the standard datasets in the field of object detection. The experiments of our proposed multi-scale feature selective matching network (MFSMNet) are also based on these.

Because the images in the PASCAL VOC 2007 datasets and PASCAL VOC 2012 datasets are mutually exclusive, numerous object detection techniques combine training on the PASCAL VOC 2007 and 2012 datasets with evaluation on images from the PASCAL VOC 2007 evaluation set. After merging, there are 16,551 training images with 40,058 target objects and 4952 evaluation images with 12,032 objects. For the evaluation metrics of the model on PASCAL VOC 2007 + 2012 datasets, we use the mean average precision (mAP).

Because the Microsoft COCO 2017 datasets have more images and objects than Microsoft COCO 2014, which makes the Microsoft COCO 2017 datasets more challenging, we select the Microsoft COCO 2017 datasets. The Microsoft COCO 2017 datasets have over 118,000 training set images, 910,670 target annotations, and 5000 evaluation set images. We also use Microsoft COCO evaluation criteria, such as: average precision (AP), average precision of small-size objects (AP_S), average precision of medium-size objects (AP_M), and average precision of large-size objects (AP_L).

4.2. Experimental Setup

Our proposed multi-scale feature selective matching network (MFSMNet) for object detection is implemented through MMDetection [41], a toolbox for object detection based on Pytorch. In ablation experiments, quantitative experiments, and qualitative experiments, we trained and predicted using 1 GeForce RTX 3090. The experimental parameters of our proposed MFSMNet method on the PASCAL VOC2007 + 2012 datasets are set as follows: the backbone is ResNet-50; the image input size is 1000×600; the optimizer is SGD; the learning rate is 2×10^{-2}; and the weight decay is 10^{-4}. The method is trained for 12 epochs. With other comparative representative methods, the batch size is set to 16. The experimental parameters of our proposed MFSMNet method on the Microsoft COCO 2017 datasets are set as follows: the backbone is ResNet-50; the maximum size of the input image is rescaled to 1333×800 without changing the aspect ratio; the optimizer is SGD; the learning rate is 2×10^{-2}; and weight decay is 10^{-4}. The method is trained with other comparative representative methods for 12 epochs, and the batch size is set to 8.

4.3. Quantitative Analysis of Ablation Experiments

Our ablation experiments are based on the Microsoft COCO 2017 datasets with a ResNet-50 backbone, using a 12 epoch training scheme. In addition, the maximum size of the

input image is rescaled to 1333 × 800 without changing the aspect ratio. Further, to show the efficacy of our approach, we conduct ablation experiments on the MSEM structure and the ASM strategy. In order to accomplish this, we remove the MSEM structure and ASM strategy from our suggested MFSMNet. The resulting network is then utilized as a baseline for comparisons.

Our proposed MFFMNet method consists of a multi-scale semantic enhancement module (MSEM) and anchor selective matching (ASM) strategy. To verify that the MSE structure can enhance multi-scale semantic information and thus improve the detection performance of small-size targets, the experimental results of Baseline + MSEM are used to demonstrate the effectiveness of the MSEM structure. With the Baseline + MSEM structure, the AP_S improved from 21.6% to 22.7%, with a performance gain of 1.1%. This demonstrates the effectiveness of the MSEM structure for small-size targets. In addition, the AP of Baseline + MSEM improved from 38.6% to 38.9%, and this average accuracy improvement was not significant. The reason for this is that the MSEM structure is designed to mitigate the feature loss of small-sized targets during down-sampling, so the AP improvement is not significant.

In order to show that the ASM strategy can anchor selection matching and thus improve detection performance, the effectiveness of the ASM strategy is demonstrated by the experimental results of Baseline + ASM. With the Baseline + ASM strategy, the AP reached 39.6%, and the AP_S, AP_M, and AP_L reached 22.2%, 43.8%, and 51.4%, respectively. Compared with Baseline, the AP of Baseline + ASM structure improved by 1%, and the AP_S, AP_M, and AP_L improved by 0.6%, 1.4%, and 2.4%, respectively. The reason is that our proposed ASM strategy addresses the problem of Intersection over Union (IoU) as a threshold to limit the detection performance of the detector, and ASM is a novel anchor selective matching strategy, because the anchor selection and matching is based on a multi-scale anchor.

Finally, the AP of MFSMNet using Baseline + MSEM + ASM structure reached 39.9%, and the AP_S, AP_M, and AP_L reached 23.2%, 44.2%, and 51.9%, respectively. Compared with the Baseline, the AP of MFSMNet with Baseline + MSEM + ASM structure is improved by 1.3%, whereas the AP_S, AP_M, and AP_L are improved by 1.6%, 1.8%, and 2.9%, respectively. The ablation experiments demonstrated the effectiveness of MFSMNet, especially for small object detection. The results of the quantitative experiments are shown in Table 1 (the bold font in the table is the highest detection accuracy of the category).

Table 1. Quantitative results of ablation experiment.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	38.6	56.4	41.7	21.6	42.4	49.0
+MSEM	38.9	56.5	41.9	22.7	42.4	49.4
+ASM	39.6	56.9	43.2	22.2	43.8	51.4
Baseline + MSEM + ASM	**39.9**	**57.1**	**43.6**	**23.2**	**44.2**	**51.9**

4.4. Qualitative Analysis of Ablation Experiments

In order to demonstrate the detection performance of our proposed MSEM structure and ASM strategy and to verify the quantitative experimental results, we perform qualitative experiments and analysis of the ablation experiments. To more objectively demonstrate the effectiveness of the MSEM structure and ASM strategy, this experiment uses the Microsoft COCO 2017 datasets for network training and the images from the PASCAL VOC 2007 + 2012 datasets for inference. In order to better reflect the detection effect of small-size targets, example images rich in small-size targets are selected from the datasets. Otherwise, each parameter setting is the same as above.

As shown in Figure 3, Figure 3a shows the visual detection results of Baseline, where there is a significant object miss in the phone booth in the middle of the image. In addition, the confidence level of the detected targets is not satisfactory; for example, the confidence level of the person in the lower left corner of the image is only 79%. In Figure 3b, the visualization detection results using the Baseline + MSEM structure are shown, and the visual observation does not reveal the existence of obvious missed objects. In addition, the photo of the person in the phone booth in the middle of the image is no longer missed, the confidence level of each target is significantly improved, and the small-sized targets in the distance are also detected better. However, there is an obvious false detection frame in the image (misdetection of a bench under the phone booth). This is due to the fact that Figure 3b uses only the Baseline + MSEM structure to further enhance the semantic information in the feature map without a corresponding sample assignment strategy. Figure 3c using Baseline + ASM strategy compared with Figure 3a using Baseline only; the target miss detection is effectively improved and there is no obvious object miss detection. At the same time, the confidence level of each target frame is significantly improved, and the confidence level converges to a reasonable interval range. The visualization results using the Baseline + MSEM + ASM strategy are shown in Figure 3d, where the confidence level of the target frame has been further improved, and there are no significant misses and false detections. In addition, the human-shaped display panel in the phone booth in the middle of the image gives a more reasonable confidence level. In addition, the detection performance of small target detection has also been more obviously improved. Taking the handbag in the middle of the image as an example, the near handbag has a high confidence level, whereas the distant handbag lacking semantic information also achieves the detection effect. In general, after using Baseline + MSEM + ASM strategy, our proposed MFSMNet using ASM strategy can enhance the MSEM structure with multi-scale semantic information for more effective anchor selection, thus achieving better detection results, especially for small objects.

Figure 3. Visualization of ablation results: (**a**) Baseline, (**b**) Baseline + MSEM, (**c**) Baseline + ASM, (**d**) Baseline + MSEM + ASM.

4.5. Quantitative Analysis of Comparative Experiments

To demonstrate the competitiveness of our proposed MFSMNet, in this subsection MFSMNet is compared with other 10 representative methods (RetinaNet [35], FSAF [42], Reppoints [15], FCOS [22], ATSS [43], Foveabox [44], GFL [45], VFNet [46], Free Anchor [47], and YOLOv5-s) for quantitative experiments on the PASCAL VOC 2007 + 2012 datasets and the Microsoft COCO 2017 datasets.

The results of the quantitative experiments on the PASCAL VOC 2007 + 2012 datasets are shown in Table 2 (the bold font in the table is the highest detection accuracy for this category). The MFSMNet method achieves the best experimental results in terms of mAP compared to the other 10 representative methods. In addition, the highest accuracy was achieved for seven categories on the PASCAL VOC 2007 + 2012 datasets.

The results of the quantitative experiments on the Microsoft COCO 2017 datasets are shown in Table 3 (the bold font in the table is the highest detection accuracy for this category). Our proposed MFSMNet method achieves the best experimental results in all six metrics compared with nine other representative methods. From the experimental results, it was shown that it reached 39.9% in AP, which is only 0.3% higher compared to GFL [45], but 1.4% higher in AP_S, proving that MFSMNet has merit in small object detection. Compared with the anchor assignment-based target detection methods ATSS [43] and Free Anchor [47] , our proposed MFSMNet is 1.3% and 1.7% higher in AP, respectively. Meanwhile, in terms of the AP_S of small objects, MFSMNet is 1.6% and 2.4% higher than the other two methods based on anchor assignment-based methods, respectively. In addition, our proposed MFSMNet has the same advantage in detection accuracy for medium and large objects.

4.6. Qualitative Analysis of Comparative Experiments

In order to demonstrate the detection performance of our proposed MFSMNet method, we select the top two quantitative results (Table 3) for visualization. The first is our proposed MFSMNet, and the second is GFL [45] for visualizing the results. To ensure that the visualization results actually reflect the real performance of the model, the images used for the visualization result inference in this subsection are from the PASCAL VOC 2007 + 2012 datasets, whereas the training set of the model is from the Microsoft COCO 2017 datasets. In the process of inference for the two methods, the parameters are set as above.

In Figure 4, to demonstrate the advantages of our proposed MFSMNet for multi-scale detection, images with multi-scale targets are selected. The first column is the input image of the detection network, the second column is the detection result of our proposed MFSMNet, and the third column is GFL [45] (second place of detection accuracy in Table 2 of Microsoft COCO 2017 datasets). In the first row of horse racing images, our proposed MFSMNet has a significant advantage in the localization of bounding boxes as well as in the category confidence. In the second row of the image of the sailboat, the detection result of GFL [45] in the lower right corner of the image shows a false detection for the surfboard, whereas our proposed object detection method of MFSMNet achieves a correct detection using local information. In the image of the third row of carriages, our proposed MFSMNet has obvious advantages in the detection of people, although the detection of horses is slightly lacking. In the last row of detection of small object bikes, our proposed MFSMNet has better performance than GFL [45], which misdetects small object bikes as motorcycles in the detection of small-size objects. In summary, the visualization detection results verify the quantitative experimental results and verify the competitiveness of our proposed MFSMNet.

Table 2. Quantitative experimental results of PASCAL VOC 2007 + 2012 datasets.

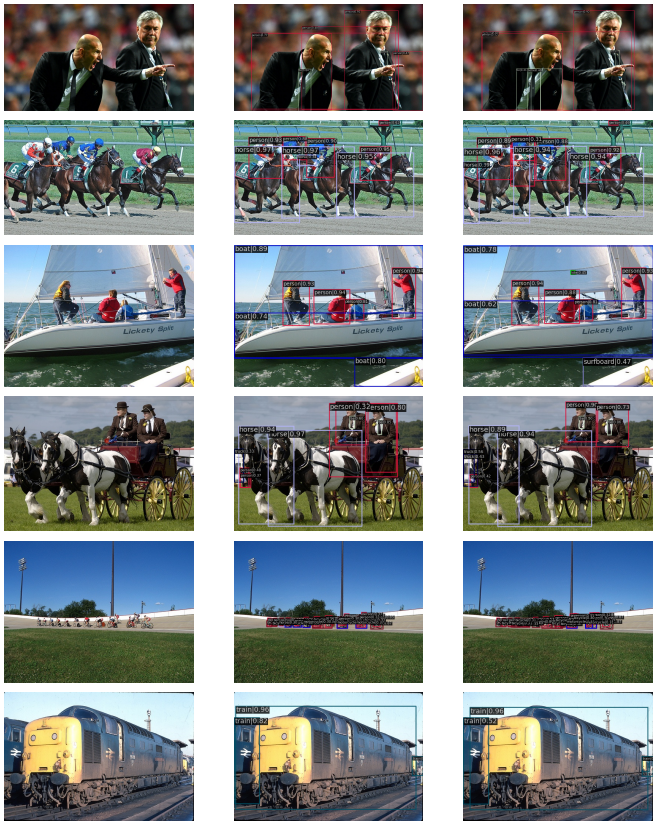
Method	*mAP*	Plane	Bike	Bird	Boat	Cup	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Human	Plant	Sheep	Sofa	Train	TV
RetinaNet [35]	79.13	87.2	86.2	78.1	66.4	71.0	84.7	87.7	88.4	62.9	85.2	72.6	85.8	85.7	82.7	84.1	53.4	82.9	**77.0**	82.9	77.7
FSAF [42]	76.31	79.3	79.3	76.0	65.1	67.6	83.1	86.7	87.1	59.8	83.2	69.5	85.3	85.1	81.9	84.4	48.2	76.3	71.6	82.8	74.0
Repponits [15]	79.47	83.5	82.4	77.1	**72.4**	**71.6**	85.1	87.8	88.3	**63.4**	86.3	75.7	87.5	85.8	84.1	83.8	50.7	**84.0**	76.2	86.3	77.4
FCOS [22]	71.59	78.5	78.7	68.3	61.8	57.6	78.0	82.2	83.0	54.8	80.2	65.8	80.4	78.4	77.4	76.5	41.4	74.5	66.9	81.1	66.2
ATSS [43]	77.77	84.7	81.9	76.8	67.9	69.5	85.4	86.4	88.1	61.7	**86.4**	72.3	85.1	85.4	80.2	83.1	48.3	81.1	72.3	81.7	77.1
Foveabox [44]	76.67	79.8	80.2	77.0	66.9	66.7	82.5	86.9	87.3	62.1	85.6	69.4	85.1	85.9	78.9	84.4	48.8	79.1	71.2	79.4	76.5
GFL [45]	77.04	85.4	83.6	76.1	63.9	67.6	82.2	86.5	86.9	59.6	83.4	72.8	83.9	84.9	83.2	83.3	48.7	78.2	70.6	83.9	76.2
VFNet [46]	77.83	83.1	84.3	76.7	68.4	69.5	84.5	86.7	87.3	61.3	83.7	70.5	84.8	85.4	83.4	84.2	49.8	79.0	73.1	84.3	76.6
Free Anchor [47]	78.16	85.0	83.6	76.0	65.5	69.7	85.4	86.9	87.6	62.5	82.3	72.3	85.1	86.0	84.4	85.0	47.5	82.3	74.9	**85.1**	76.1
YOLOv5-s [1]	77.30	87.2	**87.4**	71.8	66.4	68.6	**86.2**	**91.0**	83.5	52.0	81.3	72.2	80.0	**86.9**	**85.8**	85.3	51.1	80.3	67.8	82.5	**79.0**
Ours	**79.73**	**87.3**	85.6	**79.0**	70.8	70.9	85.7	87.4	**89.0**	63.2	85.9	**73.4**	**85.9**	86.7	84.3	**85.4**	**53.8**	82.4	75.5	84.4	78.0

[1] YOLOv5. [Online]. Available: https://github.com/ultralytics/yolov5 (accessed on 15 April 2023).

Table 3. Quantitative experimental results of Microsoft COCO 2017 datasets.

Method	Backbone	Input Shape	Params	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [35]	ResNet-50	3, 1280, 800	37.7	25.5	36.2	55.1	38.7	20.4	39.8	46.8
FSAF [42]	ResNet-50	3, 1280, 800	35.3	28.3	37.0	56.2	39.4	20.3	40.1	47.8
Repponits [15]	ResNet-50	3, 1280, 800	36.6	27.1	37.4	56.8	40.3	21.9	41.4	48.3
FCOS [22]	ResNet-50	3, 1280, 800	32.0	19.6	36.9	45.8	39.3	20.7	40.1	47.2
ATSS [43]	ResNet-50	3, 1280, 800	32.1	28.3	38.6	56.4	41.7	21.6	42.4	49.0
Foveabox [44]	ResNet-50	3, 1280, 800	36.2	29.6	35.5	54.9	37.8	19.8	39.1	46.1
GFL [45]	ResNet-50	3, 1280, 800	32.2	28.4	39.6	57.3	42.7	21.8	43.5	51.8
VFNet [46]	ResNet-50	3, 1280, 800	32.7	26.3	37.5	53.9	40.5	21.0	41.0	49.0
Free Anchor [47]	ResNet-50	3, 1280, 800	38.3	25.3	38.2	56.7	40.7	20.8	41.6	49.8
YOLOv5-s [1]	Draknet-53	3, 640, 640	**7.2**	**140.8**	37.1	57.0	39.6	20.9	42.6	47.6
Ours	ResNet-50	3, 1280, 800	32.9	22.8	**39.9**	**57.1**	**43.6**	**23.2**	**44.2**	**51.9**

[1] YOLOv5. [Online]. Available: https://github.com/ultralytics/yolov5 (accessed on 15 April 2023).



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

Figure 4. Comparison of the visual detection results of our proposed MFSMNet with the second method on the PASCAL VOC datasets: (**a**) input image, (**b**) ours, (**c**) GFL [45].

5. Conclusions

In this paper, we propose the multi-scale feature selection matching network (MFSMNet) for object detection. First, we construct the multi-scale semantic enhancement module (MSEM), which can obtain richer semantic information of small-size objects, and thus improve the detection accuracy of small-size objects. After that, we design the anchor selective matching strategy (ASM) strategy, which alleviates the positive and negative sample imbalance problem and improves the performance of multi-scale object detection. Finally, MFSMNet has advantages in detection accuracy compared with 10 other representative object detection methods, and our proposed MFSMNet can achieve the highest detection

accuracy among the 11 methods. In particular, the detection performance of MFSMNet has obvious performance improvement for small-size objects detection.

Author Contributions: Y.P.: methodology, writing—original draft preparation, experiments. Y.D.: conceptualization, methodology, writing—reviewing and editing. L.Z.: methodology, investigation, experiments. J.M.: methodology, investigation, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Henan under Grant 232300421023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data generated and analyzed during this study are available from the corresponding author by request.

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

References

1. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
2. Zhou, D.; Liu, Z.; Wang, J.; Wang, L.; Hu, T.; Ding, E.; Wang, J. Human-object interaction detection via disentangled transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19568–19577.
3. Lee, S.J.; Lee, S.; Cho, S.I.; Kang, S.J. Object detection-based video retargeting with spatial–temporal consistency. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4434–4439. [CrossRef]
4. Li, Z.; Lang, C.; Liang, L.; Zhao, J.; Feng, S.; Hou, Q.; Feng, J. Dense attentive feature enhancement for salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 8128–8141. [CrossRef]
5. 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*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
6. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
7. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
8. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
9. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
10. Dong, Y.; Tan, W.; Tao, D.; Zheng, L.; Li, X. CartoonLossGAN: Learning surface and coloring of images for cartoonization. *IEEE Trans. Image Process.* **2021**, *31*, 485–498. [CrossRef] [PubMed]
11. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
12. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 734–750.
13. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.
14. Maninis, K.K.; Caelles, S.; Pont-Tuset, J.; Van Gool, L. Deep extreme cut: From extreme points to object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 616–625.
15. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9657–9666.
16. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
17. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
18. Dong, Z.; Li, G.; Liao, Y.; Wang, F.; Ren, P.; Qian, C. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10519–10528.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. Densebox: Unifying landmark localization with end to end object detection. *arXiv* **2015**, arXiv:1509.04874.

21. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. 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.
22. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Repulic of Korea, 27 October–2 November 2019; pp. 9627–9636.
23. Dong, Y.; Jiang, Z.; Tao, F.; Fu, Z. Multiple spatial residual network for object detection. *Complex Intell. Syst.* **2022**, *9* , 1347–1362. [CrossRef]
24. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
25. Girshick, R. Fast r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28* . [CrossRef] [PubMed]
27. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2021; pp. 3520–3529.
28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
29. Lee, Y.; Park, J. Centermask: Real-time anchor-free instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13906–13915.
30. Dong, Y.; Zhao, K.; Zheng, L.; Yang, H.; Liu, Q.; Pei, Y. Refinement Co-supervision network for real-time semantic segmentation. *IET Comput. Vis.* **2023**, in press. [CrossRef]
31. Dong, Y.; Yang, H.; Pei, Y.; Shen, L.; Zheng, L.; Peiluan, L. Compact interactive dual-branch network for real-time semantic segmentation. *Complex Intell. Syst.* **2023**, in press. [CrossRef]
32. Fang, F.; Li, L.; Zhu, H.; Lim, J.H. Combining faster R-CNN and model-driven clustering for elongated object detection. *IEEE Trans. Image Process.* **2019**, *29*, 2052–2065. [CrossRef] [PubMed]
33. Dong, Y.; Shen, L.; Pei, Y.; Yang, H.; Li, X. Field-matching attention network for object detection. *Neurocomputing* **2023**, *535*, 123–133. [CrossRef]
34. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
35. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
36. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Beijing, China, 17–20 September 2017; pp. 4203–4212.
37. Fang, F.; Xu, Q.; Li, L.; Gu, Y.; Lim, J.H. Detecting objects with high object region percentage. In Proceedings of the 2020 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 7173–7180.
38. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 23–27 July 2018 ; pp. 2117–2125.
39. Everingham, M.; Winn, J. *The Pascal Visual Object Classes Challenge 2007 (voc2007) Development Kit*; Tech. Rep; University of Leeds: Leeds, UK , 2007.
40. 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*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
41. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
42. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
43. 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/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
44. 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. [CrossRef]
45. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
46. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. Varifocalnet: An iou-aware dense object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8514–8523.
47. Zhang, X.; Wan, F.; Liu, C.; Ji, X.; Ye, Q. Learning to match anchors for visual object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3096–3109. [CrossRef] [PubMed]