Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.052759

ARTICLE

Two-Layer Attention Feature Pyramid Network for Small Object Detection

Sheng Xiang[1], Junhao Ma[1], Qunli Shang[1], Xianbao Wang[1,*] and Defu Chen[1,2]

[1]College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China

[2]Binjiang Cyberspace Security Institute of ZJUT, Hangzhou, 310056, China

*Corresponding Author: Xianbao Wang. Email: wxb@zjut.edu.cn

Received: 14 April 2024 Accepted: 06 June 2024 Published: 20 August 2024

ABSTRACT

Effective small object detection is crucial in various applications including urban intelligent transportation and pedestrian detection. However, small objects are difficult to detect accurately because they contain less information. Many current methods, particularly those based on Feature Pyramid Network (FPN), address this challenge by leveraging multi-scale feature fusion. However, existing FPN-based methods often suffer from inadequate feature fusion due to varying resolutions across different layers, leading to suboptimal small object detection. To address this problem, we propose the Two-layer Attention Feature Pyramid Network (TA-FPN), featuring two key modules: the Two-layer Attention Module (TAM) and the Small Object Detail Enhancement Module (SODEM). TAM uses the attention module to make the network more focused on the semantic information of the object and fuse it to the lower layer, so that each layer contains similar semantic information, to alleviate the problem of small object information being submerged due to semantic gaps between different layers. At the same time, SODEM is introduced to strengthen the local features of the object, suppress background noise, enhance the information details of the small object, and fuse the enhanced features to other feature layers to ensure that each layer is rich in small object information, to improve small object detection accuracy. Our extensive experiments on challenging datasets such as Microsoft Common Objects in Context (MS COCO) and Pattern Analysis Statistical Modelling and Computational Learning, Visual Object Classes (PASCAL VOC) demonstrate the validity of the proposed method. Experimental results show a significant improvement in small object detection accuracy compared to state-of-the-art detectors.

KEYWORDS

Small object detection; two-layer attention module; small object detail enhancement module; feature pyramid network

1 Introduction

Object detection is a core technology in computer vision, whose task is to recognize the location and class of a specific object in the image. In recent years, object detection has made great breakthroughs and has been widely used in many real-world applications, including object tracking [1,2], pedestrian monitoring [3], intelligent driving [4], and various other domains. During this period, many

object detection algorithms including Faster Region-based Convolutional Neural Network (Faster R-CNN) [5] and You Only Look Once (YOLO) [6] were proposed. However, the above object detection algorithms are aimed at general objects. In the real scene, the phenomenon of small objects often appears due to the different conditions such as the angle and distance of the image shooting. At present, the conventional object detection algorithm still has some problems such as wrong detection and missing detection when facing these situations, and the detection accuracy is not ideal.

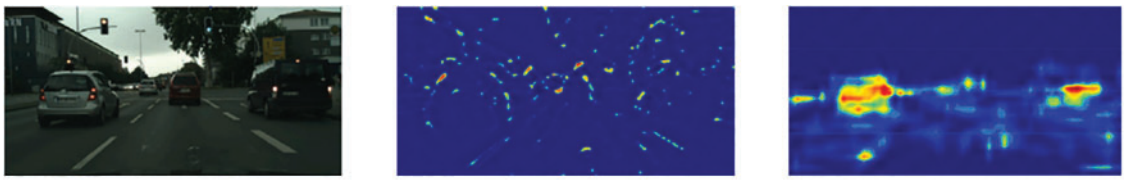
Small objects are typically defined as those measuring less than 32×32 pixels in size [7], accurate detecting small objects is a very challenging task but is used extensively. Due to their small number of effective pixels and ease of being submerged by the background, small object detection faces major challenges, such as difficulties in feature extraction, high positioning accuracy requirements, uneven samples, and dense distribution of small objects. However, accurate detection of these objects has widespread applications [8,9]. For example, small object detection can improve the safety of automatic driving, enabling vehicles to quickly and accurately identify road signs, pedestrians, bicycles, and other small obstacles, thereby ensuring better driving safety. If these small objects cannot be accurately detected, driving safety is greatly reduced, and even life safety is endangered. Therefore, in these application contexts, selecting appropriate methods to detect small objects accurately is very important.

In recent years, Convolutional Neural Networks (CNN) have been rapidly developed, and many effective methods have been proposed. Yang et al. [10] proposed a query mechanism to query potential locations of small objects on different resolution feature maps. Min et al. [11] proposed to detect small objects by emphasizing object cues and reducing redundant noise. However, these methods only capture object information at a single level, without considering the connections between the different layers of the Feature Pyramid Network (FPN). Mahaur et al. [12] added new bottom-up paths to FPN, and Huang et al. [13] introduced innovative fusion techniques to enhance information interaction between different layers and fuse multiple layers of information, but they do not make full use of the bottom layer feature of FPN, which is the key to detecting small objects. In these methods, most researchers extract multi-scale feature maps to detect small objects, among which FPN [14] is a common detection method using multi-scale feature maps.

However, although FPN improves the multi-scale performance of small objects, some problems still affect the further improvement of detection performance. As shown in Fig. 1, different layers contain different features. The 1/32 scale (high-scale) feature of the FPN emphasizes the semantic information of the object, while the neighboring 1/16 scale (low-scale) output extracts object boundaries. When these features are fused, the high-level semantic information will find relevant texture information in the lower level to be fused instead of blindly fused. However, since these lower layer features do not contain enough semantic information by themselves, but only object boundary information such as dots, lines, edges, etc., they cannot provide enough semantic guidance for the higher layer features in the fusion process. The direct fusion of these features without considering the semantic gaps between different layers will lead to mutual interference of semantic information, which is not conducive to the effective expression of multi-scale features, and small object information may be submerged in the interference information.

Moreover, as shown in Fig. 1, since CNN utilizes pooling layers and convolution layers repeatedly to extract semantic information, small objects may be lost during the down-sampling process, for example, the small object can no longer be observed in (c). Therefore, the high layers of FPN do not contain enough small object information, and the low layers, especially the bottom layer, have the most abundant small object information. Fig. 1 shows that in a common FPN detector, the direct fusion of

features across scales still reduces the small object detection accuracy. Therefore, further research on small object detection is necessary, as it can enhance the performance of the network. Therefore, we consider strengthening the close connection between different layers and enhancing the characteristics of high-level small objects to improve the representation ability of small objects, and to improve the effect of small object detection.



(a) (b) (c)

Figure 1: The motivation for this work. (a) is the original image, (b) and (c) come from the 1/16 and 1/32 scales of the ResNet backbone

Inspired by the above observations, we propose a Two-layer Attention Feature Pyramid Network (TA-FPN). Unlike previous FPN-based approaches, TA-FPN fully utilizes the information of each feature layer, deepening the connection between neighboring layers while using the bottom layer to enhance the small object features of other layers. On the one hand, we propose a Two-layer Attention Module (TAM), which utilizes the attention module to make the network more attentive to the semantic information of the objects and integrates it into the lower layers for mitigating the semantic gaps between different layers so that each layer contains similar semantic information. One advantage of TAM is that it not only focuses on the object region at a single feature layer but also fuses the feature information of two adjacent layers. On the other hand, to compensate for the insufficient information on high-level small objects caused by downsampling, a novel Small Object Detail Enhancement Module (SODEM) is adopted to strengthen local features of objects, suppress background noise, supplement information details of small objects, and fuse the enhanced features to other feature layers. This ensures each layer is rich in small object information, enhances the utilization of small objects, further strengthens the small object feature representation of each layer, and finally improves the accuracy of small object detection. This module can focus part of the attention of high-level features on small object information, which helps express small object feature information.

The attention should be drawn to the fact that TAM employs dilated convolutions with varying dilation rates across multiple branches to capture both local and global contextual information. This approach enables the extraction of semantic information at different levels, thereby better representing the comprehensive information of objects. At the same time, the module uses a local attention mechanism to make the fused features pay more attention to local features, which helps in the accurate detection of small objects. The SODEM extensively leverages the abundant small object information in the bottom layer of the FPN and inputs high-resolution features into each feature layer, allowing it to preserve details of small objects.

The main contributions of this paper can be summarized as follows:

(1) We propose a Two-layer Attention Module to extract rich semantic information and make the adjacent layers contain similar semantic information by fusing to the lower layers, so as to alleviate the semantic gaps between different layers, and at the same time make the network focus on small objects, and improve the accuracy of small object detection.

(2) We propose an efficient Small Object Detail Enhancement Module, which strengthens the local features, inhibits the background noise, enhances the information details of the small object, and fuses the enhanced features into other feature layers to ensure that each layer is rich in small object information, thus further strengthens the small object feature expression of each layer.

(3) Experimental results on MS COCO and PASCAL VOC datasets show that TA-FPN effectively improves the performance of small object detection.

2 Related Work

2.1 Generic Object Detection

Existing detectors based on CNN are primarily classified into one-stage and two-stage detectors. The typical two-stage detector [15,16] generates a region of interest (ROI) and then uses a classifier and regression to refine the ROI. Mask R-CNN [17] uses a segmentation branch with a RoIAlign layer that significantly improves detection performance. Cascade R-CNN [18] is a cascaded object detection algorithm based on Faster R-CNN, it introduces multiple cascading stages. These two-stage detectors are computationally expensive and slow to detect. To solve this problem, single-stage detectors directly utilize feature maps for detection, which effectively improves the detection speed. YOLOv3 [19] mainly changed the activation function of class prediction from Softmax to Sigmoid of logistic regression and removed the previous width and height square root. Moreover, YOLOv3 predicted a set of class probabilities for each bounding box. Modified the rules of positive and negative sample selection. YOLOv7 [20] designed several trainable bag-of-freebies, and proposed a planned model structure re-parameterization method, adjusting the number of channels so that the detector greatly improves the detection accuracy without increasing the amount of computation, but there is no appropriate sensitivity field to detect small objects. Despite the good improvement in object detection accuracy, small object detection is still an unsolved challenge, because the general object detection measurement is all scales and detectors dedicated to small objects still need more development.

2.2 Small Object Detection

Because of the small size of the small objects, it is relatively clear in high resolution, so most of the small object detection methods detect small objects on high-resolution feature maps. Bai et al. [21] proposed to super-distinguish ROI, but many ROIs still require a lot of calculation. Unlike super revolution (SR) for ROI, feature-level SR directly processes features with super-resolution, reducing the amount of computation but lacking direct supervision. Noh et al. [22] proposed a direct supervision method for small-scale objects during the training process. Bosquet et al. [23] proposed to combine an object generator based on Generative Adversarial Networks (GANs) with image restoration and image blending techniques to obtain high-quality synthetic data, which can generate small objects from large objects. Deng et al. [24] proposed to rebuild a high-resolution additional feature layer, especially for small object detection outside the original pyramid structure, but there was no effective information exchange among the feature layers. To deepen the connection between different layers, Min et al. [25] proposed to highlight low-level small goals by filtering redundant semantics and using detailed context information to detect small objects. Although the above methods have improved the performance of small object detection, they have not fully utilized the bottom feature of FPN. In our work, based on FPN, we pay more attention to the connection between different layers, especially between adjacent layers and between the bottom layer and other layers, because the bottom layer is the key to detecting small objects.

2.3 Detection with Multi-Scale Features

Utilizing multi-scale features is an effective approach to mitigate the issues caused by object scale variations, the representative method is FPN. At present, many frameworks based on FPN have been extended, which greatly improves the performance of object detection. Liu et al. [26] proposed a path aggregation network to enhance the information of the bottom layer through the bottom-up path of FPN. To balance the features of various layers, Pang et al. [27] proposed to provide equal weights to the features at each level. To make the most of each layer of information, Tan et al. [28] introduced a weighted bidirectional feature pyramid network (BiFPN), which enables fast feature fusion, uniformly scaling the resolution of all features. These FPN-based methods improve the precision of small object detection but do not focus on the semantic gaps between the feature layers in the top-down process. Inspired by them, we design a new network that alleviates the semantic gaps between different layers and provides small object information for each layer.

3 Proposed Method

To address the decrease in small object detection accuracy caused by the semantic gaps in the direct fusion of features at different scales and the loss of small objects during downsampling, we propose a Two-layer Attention Feature Pyramid Network (TA-FPN), as shown in Fig. 2. TA-FPN comprises the Two-layer Attention Module (TAM) and the Small Object Detail Enhancement Module (SODEM). On the one hand, TAM is responsible for extracting advanced semantic information, which is subsequently integrated into lower layers, making the model more focused on the object region, and mitigating the semantic gaps. On the other hand, SODEM aims to fully utilize the bottom layer information within the FPN, ensuring that each layer is enriched with abundant small object information to enhance small object detection accuracy. In this section, we introduce the proposed TA-FPN, which can be considered as an integration of FPN in Section 3.1, TAM in Section 3.2, and SODEM in Section 3.3. Section 3.4 is the loss function of TA-FPN.

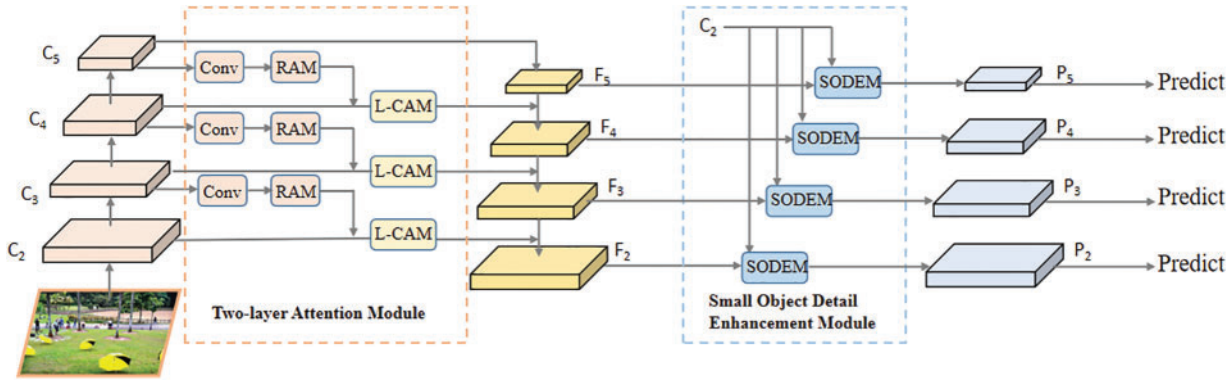


Figure 2: Overall architecture of TA-FPN. It comprises a two-layer attention module (TAM) and a small object detail enhancement module (SODEM). Conv denotes dilation convolution

3.1 Framework of Proposed Method

We mainly use ResNet [29] as BackBone, which is because the problem of gradient vanishing or gradient explosion may occur when the network depth increases, thus affecting the model accuracy, which the ResNet can address well. ResNet mainly consists of residual blocks, which are composed of multiple cascaded convolutional layers and a shortcut connection. After fusing the output values of

these two parts, the output is obtained through the Rectified Linear Unit (ReLU) activation function. To better detect small objects, we adopt the framework of FPN after ResNet extracts image features.

The low layers of FPN usually have high resolution, which is beneficial for small object localization. Additionally, the high layers obtain more semantic information but the spatial resolution is compromised. However, FPN fails to account for the semantic gaps between different layers, and direct fusing these features causes the semantic information to interfere with each other, and a large number of small object information is submerged in the interference information. A lot of small object information is lost in downsampling, which decreases the detection accuracy of small objects.

To address these problems, we proposed TA-FPN. Fig. 2 provides an overview of our algorithm. On the one hand, the Two-layer Attention Module is designed to accentuate crucial areas within the image, suppress noise in regions like the background, and bolster feature representation for small objects. Additionally, it enhances the connectivity between different layers, mitigating semantic gaps that may exist between these layers. On the other hand, the Small Object Detail Enhancement Module leverages the bottom layer of the FPN to offer a wealth of information related to small objects to the other layers, thereby enhancing the accuracy of small object detection.

3.2 Two-Layer Attention Module

FPN directly combines features from different layers without considering the semantic gaps between them. This will lead to the generation of redundant information, thereby diminishing the expressive capacity of multi-scale features, and small objects can be easily submerged. For this problem, we propose the Two-layer Attention Module (TAM), which extracts higher-level semantic information and integrates it into the lower-level layers, ensuring that neighboring layers contain consistent semantic information, thus mitigating semantic discrepancies. TAM comprises two key components: the Residual Attention Module serves to preserve high-level semantic information, while the Local Attention Module directs the model's focus more toward small objects. We will now provide a detailed introduction of each component.

In TAM, to enhance the extraction of high-quality semantic information for integration into the lower layers, one common approach is to increase the size of the convolutional kernel. This can extract more global features to get the global semantic information. However, using a large, fixed-size convolutional kernel can pose a problem, as this can result in fixed receptive fields and high computational complexity. To overcome this problem, we employ dilation convolution with varying dilation rates. This approach enables us to acquire multi-scale semantic information without sacrificing resolution. The proposed Two-layer Attention Module is shown in Fig. 3, which has two inputs, a higher-level feature $C_i \in R^{C_i \times H_i \times W_i}$ and a lower-level feature $C_{i-1} \in R^{C_{i-1} \times H_{i-1} \times W_{i-1}}$.

Specifically, to extract the rich semantic information of the high-level feature C_i, it can be first passed through a three-branch convolutional block, in which each branch contains dilation convolutions with different convolution rates (such as r = 1, 3, 5), and they can extract multi-scale semantic information from different receptive fields. The large dilation convolution rate leads to a larger receptive field, which contains more contextual information and facilitates the detection of small objects. We can represent the extracted semantic information $R^*(x)$ as:

$$R^*(x) = R_1(x) + R_3(x) + R_5(x) \tag{1}$$

where R_1, R_3, R_5 respectively denote dilation convolution with dilation convolution rates of 1, 3, 5. Then the residuals are noted, and the residuals are connected so that the pre-convolutional features

are preserved. we can produce the feature map $RAM(x)$ as follows:

$$RAM(x) = B(Conv_2(R(B(Conv_1(x))))) \tag{2}$$

where B denotes the batch specification layer and R denotes the ReLU activation function. Finally, we can obtain the enhanced features C_i^*.

$$C_i^* = \sigma(R^*(C_i) + RAM(R^*(C_i))) \otimes R^*(C_i) \tag{3}$$

where σ denotes the sigmoid activation function.

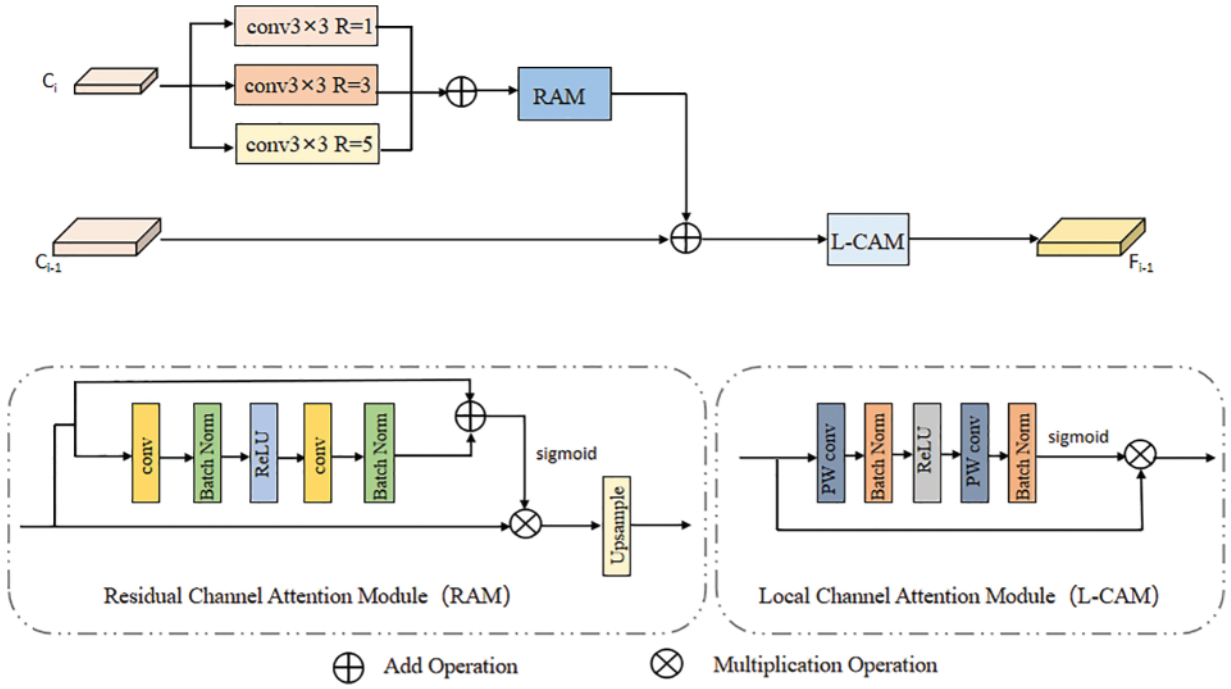


Figure 3: The framework of the two-layer attention module. The residual attention module can preserve high-level semantic information containing the main object and fuse it into lower layers to alleviate existing semantic gaps. The local attention module can make the model focus on small objects and improve the accuracy of small object detection

To enhance the small object features, we propose a local attention channel. We employ pointwise convolution (PWConv) as the context aggregator for the local channel, which only considers point channel interactions at each spatial location. After the up-sampled features have passed through the point convolution and normalization modules, we obtain an attention map that is specifically geared towards small objects as follows:

$$LCAM(x) = B(PWConv_2(R(B(PWConv_1(x))))) \tag{4}$$

where B denotes the Batch Norm layer, R denotes the ReLU activation function, and $PWConv$ denotes dot convolution. Then obtain a feature map F_{i-1} that focuses on the information of the small objects by passing the sigmoid activation function. The F_{i-1} is formulated as:

$$F_{i-1} = \sigma(LCAM(C_{i-1} + C_i^*)) \otimes (C_{i-1} + C_i^*) \tag{5}$$

where σ denotes the sigmoid activation function.

3.3 Small Object Detail Enhancement Module

Small objects usually have only a few pixels in the images, so their feature information is relatively limited and easy to confuse with the background, resulting in difficult detection. At the same time, because the common feature extraction networks usually adopt downsampling operations to reduce spatial redundancy in feature maps and obtain high-dimensional feature representations, this further exacerbates the loss of small object information. The bottom features usually contain rich local texture information, which is essential for small object detection.

Therefore, we propose the Small Object Detail Enhancement Module (SODEM) to fully utilize the bottom feature information of FPN. Specifically, the bottom feature maps that have undergone fewer downsampling operations and have higher resolutions are sequentially fused into the higher-level feature maps that are rich in semantic information but have lower resolutions. This fusion strategy can fully utilize the bottom feature maps, making the higher-level feature maps also contain the key texture information of small objects. On this basis, small object detection is performed, effectively improving the detection accuracy of the network. The introduction of SODEM enables the network to better deal with information loss, achieving performance improvement in small object detection.

The Small Object Detail Enhancement Module (SODEM) is shown in Fig. 4, where C_2 and F_i serve as inputs, and the output feature layer will contain rich small object information. Firstly, local texture information containing small object information is extracted from C_2 by using 1×1 convolution. The 1×1 convolution can highlight the local features of the object, allowing for more effective extraction of small object information. In addition, the module uses the convolutional layer to extract the local features of the objects and uses the ReLU function to help the model learn complex feature representations, retaining only the object features, which improves the clarity and recognizability of the features without being interfered by the background noise. Then, 1×1 convolution is used again to make the extracted features consistent with the number of channels in the F_i feature layer, and downsampling the extracted features.

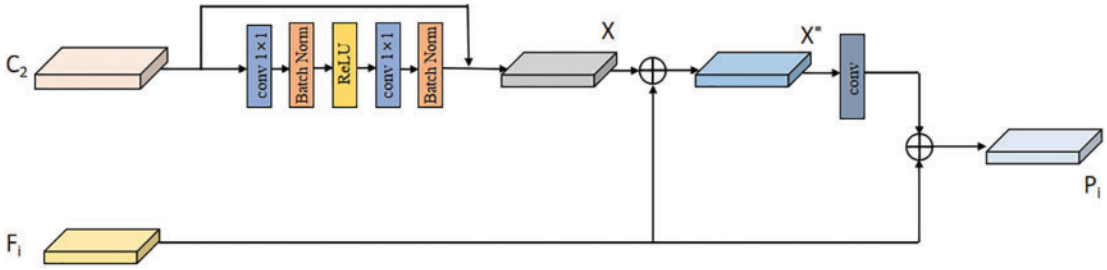


Figure 4: The framework of the small object detail enhancement module

Without this module, directly fusing the bottom features into the higher-level features would cause the noise in C_2 to be directly transmitted to the higher-level features, which could submerge small object information and affect other meaningful semantic information. This design ensures that the resulting feature map X contains both small object information and avoids noise interference. The bottom feature representation is denoted as $C_2 \in R^{C \times H \times W}$. Other feature layers are denoted as $F_i \in R^{C_i \times H_i \times W_i}$. The extracted feature map $X \in R^{C_i \times H_i \times W_i}$ can be represented as:

$$X = B(Conv(R(B(Conv(C_2))))) + C_2 \tag{6}$$

where $B(\cdot)$ denotes the Batch Norm layer and $Conv(\cdot)$ is the 1×1 convolution, $R(\cdot)$ denotes the ReLU activation function.

Next, the extracted features are fused with the F_i feature layer to ensure that this feature layer contains rich small object features, and the resulting feature map $X^* \in R^{C_i \times H_i \times W_i}$ can be expressed as:

$$X^* = X \oplus F_i \tag{7}$$

where $\oplus$ denotes element addition.

Then, a 3×3 convolution is used to minimize the influence of the aliasing effect. Finally, the output is residually concatenated with F_i, which preserves the feature information before fusion, and the output $P_i \in R^{C_i \times H_i \times W_i}$ of the SODEM module can be obtained:

$$P_i = Conv_3(X^*) \oplus F_i \tag{8}$$

where $Conv_3(\cdot)$ is a 3×3 convolution.

This module realizes the full extraction and utilization of small object features through a series of effective operations.

3.4 Loss Function

For detection, we employ the loss function:

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum p_i^* L_{reg}(t_i, t_i^*) \tag{9}$$

where L_{cls} is the classification loss function, L_{reg} is the regression loss function. N_{cls} represents the number of all samples in a mini-batch, and N_{reg} represents the number of anchor box locations. Since the difference between N_{cls} and N_{reg} is too large in practice, the weight of *cls* and *reg* is roughly equal by using the parameter λ to balance them.

(1) The classification loss function is defined as follows:

$$L_{cls}(p_i, p_i^*) = -[p_i^* log(p_i) + (1 - p_i^*) log(1 - p_i)] \tag{10}$$

where p_i is the probability of anchor i is an object. The p_i^* is 1 if the anchor is positive, and 0 otherwise.

(2) The regression loss function is the $smooth_{L1}$ loss, that is:

$$L_{reg}(t_i, t_i^*) = \sum smooth_{L1}(t_i - t_i^*) \tag{11}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if\,|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{12}$$

where t_i represents the 4 coordinates of the predicted bounding box, t_i^* represents the coordinates of the real box.

For boundary box regression, the regression parameters of the following four coordinates are used:

$$t_X = \frac{x - x_a}{w_a}, \quad t_y = \frac{y - y_a}{h_a} \tag{13}$$

$$t_W = log\left(\frac{w}{w_a}\right), \quad t_h = log\left(\frac{h}{h_a}\right) \tag{14}$$

$$t_X^* = \frac{x^* - x_a}{w_a}, \quad t_y^* = \frac{y^* - y_a}{h_a} \tag{15}$$

$$t_X^* = log\left(\frac{w^*}{w_a}\right), \quad t_x^* = log\left(\frac{h^*}{h_a}\right) \tag{16}$$

where x, y, w, and h denote the center coordinates, width, and height of the box. Variables x, x_a, and x^* are for the predicted box, anchor box, and ground-truth box, respectively.

4 Experiments

To verify the effectiveness of TA-FPN, we first performed experiments on MS COCO and Pascal VOC datasets and compared them with other state-of-the-art methods. Then, we did ablation experiments to demonstrate the effectiveness of each module proposed. The details are as follows.

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

We experiment with our method on MS COCO and PASCAL VOC.

MS COCO. MS COCO consists of more than 200 k images, and we use 115 k images for training and 5 k images for testing. MS COCO faces two main challenges in object detection: (1) Small objects: about 65% of the objects are smaller than 6% of the image size; (2) Objects with different lighting and shapes.

PASCAL VOC. We also apply our algorithm to another popular dataset, PASCAL VOC, which contains 20 different object classes and many small objects. VOC 2012 contains 11 k images and we use half for training and half for testing.

4.1.2 Evaluation Metrics

In this paper, we use the following evaluation metrics: AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L.

AP is a widely used evaluation metric. The calculation of AP is based on the precision-recall curve, which can be represented by:

$$AP = \int P(R)dR \tag{17}$$

AP_{50} and AP_{75} denote the detection accuracy when the IoU is 0.5 and 0.75, respectively. IoU can be expressed as:

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{18}$$

where B_p is the predicted box, and B_{gt} denotes the ground truth.

In addition to the AP, MS COCO can also be used to detect small, medium, and large objects, which are respectively denoted by using AP_S (Small), AP_M (Medium), and AP_L (Large) (the definitions are given in Table 1). Since we are a small object detector, we focus more on AP_S.

4.2 Implementation Details

All the experiments are implemented on PyTorch, and the GPU is an NVIDIA GTX 1080 Ti. We put the improved TA-FPN into Faster R-CNN, where ResNet is the backbone. In this experiment, SGD was used as the optimizer. During the experiment, we trained 15 epochs, the initial learning

rate was set to 0.01, and the learning rate was reduced by 1/3 after every 3 epochs. The experimental parameters setting are shown in Table 2.

Table 1: Definitions of large, medium, and small objects

	Object size (pixels)
Large	$- \geq 96 \times 96$
Medium	$32 \times 32 \leq - \leq 96 \times 96$
Small	$- \leq 32 \times 32$

Table 2: Experimental parameters

Parameters	Implication	Value
Batch size	Data batch size	4
Lr	Learning rate	0.01
Weight decay	Model weight attenuation	0.0001
Epochs	Number of training iterations	15
Optimizer	/	SGD

4.3 Comparisons with State-of-the-Arts

We conducted experiments on the COCO dataset and compared it with other detectors, and the results are shown in Table 3. When ResNet-50 and ResNet-101 as BackBone, Faster R-CNN with TA-FPN achieves an accuracy of 38.5% and 40.4%, and when using the more powerful feature extractor ResNeXt-101, our network achieves an accuracy of 42.6%. In addition, as can be seen in the AP_S, AP_M, and AP_L columns, the method also shows a substantial improvement in accuracy for small objects when compared to the Faster R-CNN base network, and achieves the best results when compared to other detectors. The medium objects, which are also slightly smaller in size, also outperform all other detectors, and the method does not give optimal results for large objects mainly because the network focuses more on localized information which may lead to the loss of some global information, nevertheless, the detection accuracy is a significant improvement. Furthermore, our method not only outperforms variants of FPN for small object detection (e.g., Path aggregation network (PANet) [26], Attentional feature pyramid network (AFPN) [25], Channel enhancement feature pyramid network (CEFPN) [30], and Hierarchical activation network (HANet) [31]), but also outperforms state-of-the-art multiscale detection methods (e.g., Graph feature pyramid network (GraphFPN) [32], Extended feature pyramid network (EFPN) [24] and Multiple spatial residual network (MSRNet) [33]).

Table 3: Comparison with state-of-the-art methods on COCO dataset

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SSD512 [34] (2016)	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
RefineDet512 [35] (2017)	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4

(Continued)

Table 3 (continued)

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet800 [36] (2017)	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
YOLOv3 [19] (2018)	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
Cascade R-CNN [18] (2018)	ResNet-101	42.8	62.1	46.3	23.7	45.5	**55.2**
PANet [26] (2018)	ResNet-50	37.5	58.6	40.8	21.5	41.0	48.6
FCOS [37] (2019)	ResNet-50	36.6	56.0	38.8	21.0	40.6	47.0
Libra R-CNN [27] (2019)	ResNet-50	38.6	60.6	42.0	22.4	41.3	47.7
Libra R-CNN [27] (2019)	ResNeXt-101	**43.0**	64.2	46.9	**25.2**	45.9	54.1
AugFPN [38] (2020)	ResNet-50	38.8	61.5	42.0	23.3	42.1	47.7
AugFPN [38] (2020)	ResNet-101	40.6	63.3	44.0	24.2	44.1	51.0
GraphFPN [32] (2021)	ResNet-101	42.1	61.3	46.1	23.6	41.1	53.3
EFPN [24] (2022)	ResNet-50	38.2	/	/	22.7	41.0	49.4
AFPN [25] (2022)	ResNet-50	38.5	61.1	41.9	22.0	42.6	49.2
AFPN [25] (2022)	ResNet-101	40.2	62.5	43.6	24.2	44.3	52.0
CEFPN [30] (2022)	ResNet-50	38.8	60.5	41.9	22.5	41.7	48.1
CEFPN [30] (2022)	ResNet-101	40.9	62.5	44.4	23.5	44.2	51.4
MSRNet [33] (2023)	ResNet-50	38.6	60.6	42.4	21.9	43.1	54.1
HANet [31] (2024)	ResNet-50	39.6	62.1	43.6	22.3	41.6	50.3
Faster R-CNN w/FPN [5]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN w/TA-FPN	ResNet-50	38.5	60.3	41.7	22.5	41.6	47.5
Faster R-CNN w/TA-FPN	ResNet-101	40.4	61.9	44.2	23.3	44.0	51.9
Faster R-CNN w/TA-FPN	ResNeXt-101	42.6	**64.6**	**47.0**	**25.2**	**46.0**	53.8

Note: If not otherwise noted in this section, the bolded text indicates the optimal outcome.

Fig. 5 shows some examples of detection results in MS COCO dataset. FPN occasionally misses some objects, such as some small objects. In contrast, the performance of TA-FPN is improved and it is able to detect more objects, especially small objects, compared to the FPN baseline. For example, in the first row, (a) is the original image, and (b) is the result detected by the FPN baseline, it can be found that the people in the near distance (large objects) can be detected, while the people and boats in the far distance (small objects) are not detected. (c) is the result detected by the TA-FPN, which not only detects people in the near distance but also detects people and boats in the far distance. In addition, to clearly show the performance of TA-FPN, (d) shows the difference between the FPN and TA-FPN detection results, which clearly shows that TA-FPN has better detection ability for small objects. Meanwhile, TA-FPN is also more robust in the face of object appearance changes or being occluded. For example, the second row (a) is the original image, and (b) is the result detected by the FPN baseline, which can only detect a person with complete object features due to the FPN alone can not focus on small object features. (c) is the result detected by TA-FPN, due to the introduction of TAM to make the model more focused on small objects and SODEM to make small object features richer, TA-FPN can also detect people with small and incomplete object features due to occlusion. Meanwhile, TA-FPN also has a better detection effect when facing environmental changes, as in the result of the third row, it can still detect people with unclear features in the distance.

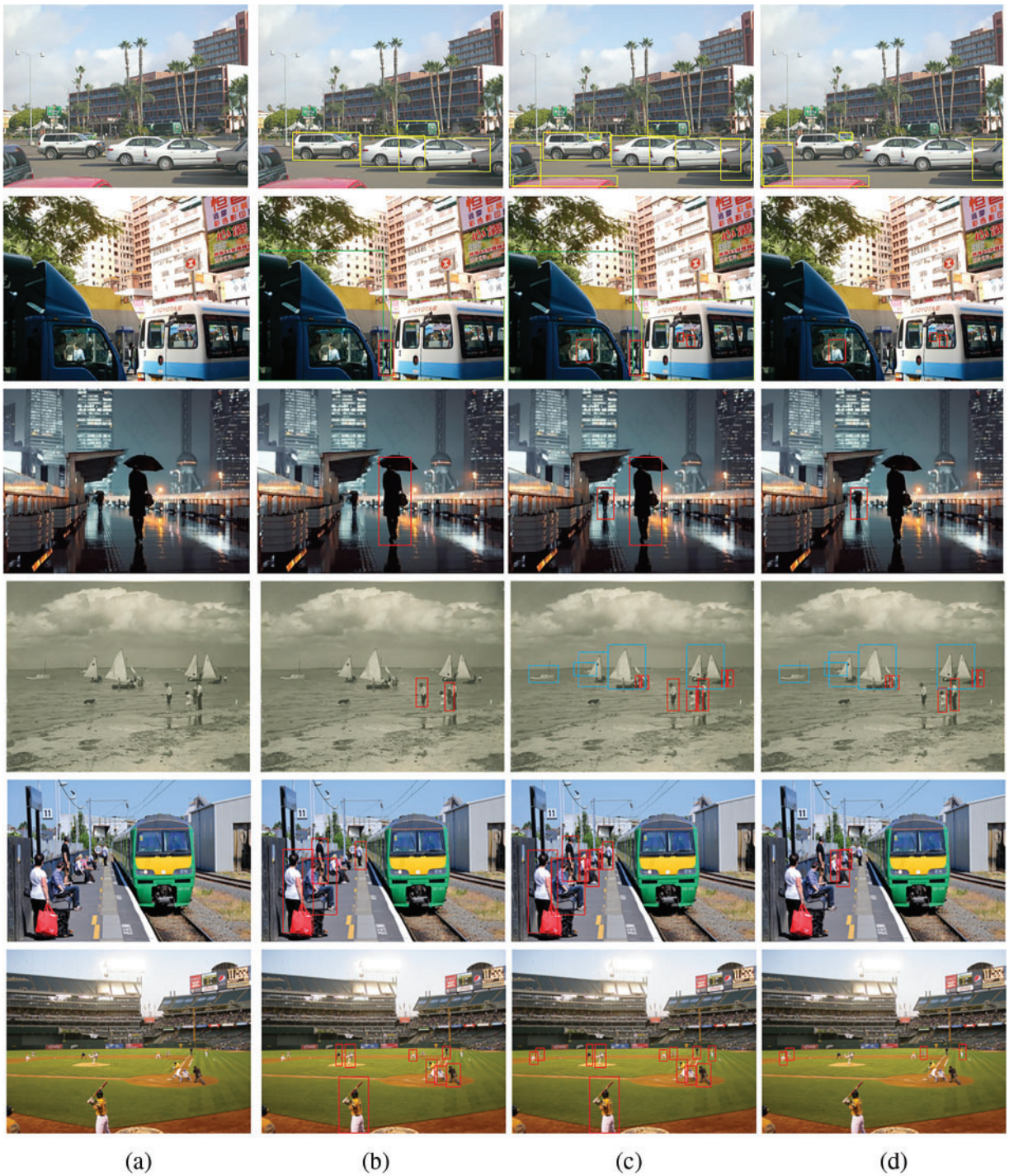


Figure 5: Detection results of proposed TA-FPN: (a) is the original image; (b) is the detection result after FPN; (c) is the result after TA-FPN; and (d) is the objects of TA-FPN detection more than FPN. Different colors represent different categories

Meanwhile, we compare the accuracy of TA-FPN with the FPN baseline in the PASCAL VOC dataset. Using ResNet-50 as BackBone, TA-FPN improves the accuracy of small object detection from 20.4% to 22.1%, and AP from 78.3% to 79.1%, which is improved 1.7% and 0.8%, respectively, which proves the effectiveness of TA-FPN.

4.4 Ablation Studies

We also analyzed the effect of each module of TA-FPN on the PASCAL VOC dataset. We conducted experiments on Faster R-CNN with ResNet-50 as BackBone and gradually added the Two-layer Attention Module and the Small Object Detail Enhancement Module. The overall experiments are shown in Table 4. The results show that both TAM and SODEM alone can improve the small object detection accuracy, and when they act synergistically, the improvement effect is more significant. The specific experimental results are analyzed as follows:

Table 4: The effectiveness of our proposed TAM and SODEM on the PASCAL VOC dataset

TAM	SODEM	mAP	AP_S	FPS
		78.3	20.4	**10.5**
√		78.8	21.0	9.1
	√	78.6	21.7	9.8
√	√	**79.1**	**22.1**	8.9

Effect of the Two-Layer Attention Module. There are semantic gaps between different layers of the FPN, and direct fusion of these features without considering the semantic gaps will generate much redundant information, and small objects can easily be drowned in noise. TAM can generate accurate semantic information from the higher layers to be passed to the lower layers so that the neighboring layers contain similar semantic information. Table 4 shows that the Two-layer Attention Module improves AP_S on the baseline by 0.6% and the average accuracy mAP by 0.5%. This suggests that the method facilitates the highlighting of possible object regions and mitigates the semantic dilution caused by the top-down process of the feature pyramid network and the semantic gaps that exist between different layers.

Effect of the Small Object Detail Enhancement Module. Small object features are often sparse and vulnerable to being submerged in noise. Common feature extraction networks can exacerbate the loss of small object features, primarily due to downsampling. This leads to lower accuracy in small object detection. The Small Object Detail Enhancement Module leverages the rich small object information present in the bottom layer of FPN. It accurately extracts detailed small object features and injects them into each layer of the network to enhance small object detection accuracy. As Table 4 illustrates, the Small Object Detail Enhancement Module improves small object detection accuracy over the baseline by 1.3%. This enhancement underscores the module's capacity to enrich small object features across the network's layers.

Effect of Different Dilation Convolution Rates. To further analyze the effect of different dilation convolution rates on the detection results, we conducted experiments to illustrate this, as shown in Table 5. To accurately detect small objects, we used a 3×3 convolutional and set the dilation convolution rate from 1 to 5 to adjust the size of the receptive field. Since we use a three-branch convolutional block, we chose three dilation convolution rates. The results show that the best results can be achieved when the dilation convolution rate is 1, 3, and 5. The possible reason for this is that the

dilation rate settings of 1, 3, and 5 can cover a larger number of objects in the sensory field, whereas settings of 1, 2, and 3 (or 3, 4, and 5) result in information redundancy because they only cover a larger amount of local information (or a larger amount of global information).

Table 5: The effectiveness of different dilation rates on the two-layer attention module (TAM)

Method	Dilations					*mAP*	AP_S
	1	2	3	4	5		
Baseline						78.3	20.4
TAM	√	√	√			78.5	20.8
	√		√		√	**78.8**	**21.0**
			√	√	√	78.5	20.7

Effect of Fusing Different Layers. To further assess which layer of the Small Object Detail Enhancement Module is fused to have the greatest impact on small object detection, we conducted experiments to illustrate this, as shown in Table 6. The table reveals that accuracy is consistently improved for each layer fusion, with the most significant improvement observed when fusing the F3 and F4 layers of FPN. The likely reason for this is that these two layers initially contain less information related to small objects, and this module effectively supplements them with rich small object information, leading to a significant improvement in accuracy. Conversely, when fusing the F5 layer, the effect is less satisfactory due to excessive downsampling. The effect improvement is also not significant when fusing the F2 layer which produces more redundant information. Notably, involving all layers in the fusion results in the most substantial improvement in small object accuracy.

Table 6: The effectiveness of fusing different layers on the small object detail enhancement module

Method	Layers				*mAP*	AP_S
	F2	F3	F4	F5		
Baseline					78.3	20.4
SODEM	√				**78.6**	20.9
		√			78.4	21.5
			√		78.5	21.3
				√	78.5	20.5
	√	√			**78.6**	20.9
	√		√		78.5	21.2
	√			√	78.4	21.0
		√	√		78.5	21.4
		√		√	78.5	21.2
			√	√	78.5	21.0
	√	√	√		78.5	20.5
		√	√	√	**78.6**	21.2
	√	√	√	√	**78.6**	**21.7**

4.5 Experimental Results Analysis

To provide a more intuitive visualization of the effectiveness of TA-FPN, we have visualized some of the features demonstrating the roles of the Two-layer Attention Module and the Small Object Detail Enhancement Module. As shown in Fig. 6, (a) is the original image, (b) is the result after normal FPN, (c) is the result after adding TAM, and (d) is the result after TAM and SODEM.

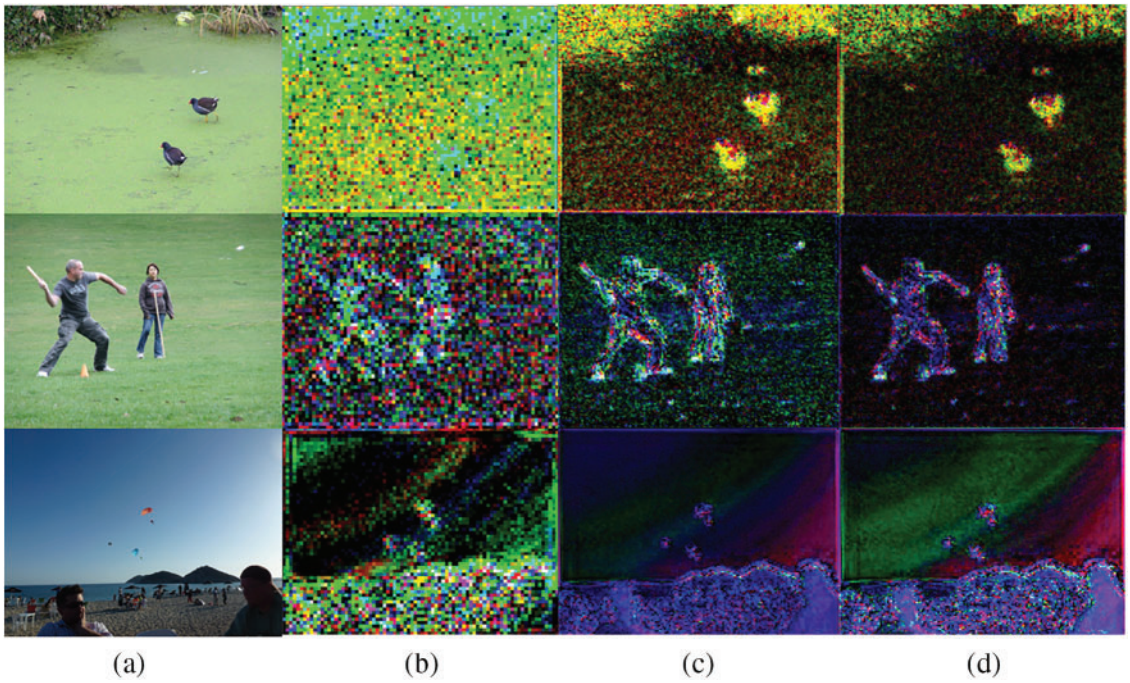


Figure 6: Experimental results of the feature maps: (a) is the original image; (b) is the feature map after FPN; (c) is the feature map with the addition of TAM; (d) is the feature map after TAM and SODEM

Since this is a small object detector, we focus on the small objects in the image. Specifically, we observe the small ball in the upper right corner of the second image, which cannot be effectively observed from (b) due to too much noise. As shown in (c), the semantic gap between different layers is mitigated due to the addition of missing information, which prevents objects, especially small objects, from being submerged in conflicts when different layers are fused. And using the local attention makes the network pay more attention to the small object area, so we can see the approximate location of the small ball, but there are still some noise and residual shadows affecting the specific localization of the small ball. As shown in (d), since SODEM supplements the small object information, it makes the small object more prominent, and without the interference of noise and residual shadows, the specific position of the ball can be observed and localized. This experiment effectively highlights the influence of each module on the detection process.

4.6 Time Complexity

We tested the time complexity of TA-FPN as shown in Table 4. When using ResNet-50 backbone, the Faster R-CNN with TA-FPN can reach 8.9 FPS and the Faster R-CNN with FPN can reach 10.5 FPS. The computing cost of our method is increased by about 15.2%. The inference speeds of TAM and SODEM are 9.1 FPS and 9.8 FPS, respectively, and the individual action of each module slightly increases the computing cost. The experimental results show that TAM and SODEM

significantly improve the performance of small object detection, but only increase a small portion of the computational resources.

5 Conclusions

In this work, we proposed a feature pyramid-based architecture called the Two-layer Attention Feature Pyramid Network, which comprises two integral components: a Two-layer Attention Module and a Small Object Detail Enhancement Module. TAM reduces the semantic gaps between different layers by implementing two-layer fusion, making the features fully fused, and highlighting the small object region to improve detection accuracy. SODEM maximizes the utility of FPN's features, especially the bottom layer feature, and fuses the rich small object information into other feature layers to strengthen the small object feature and improve the detection accuracy of small objects. Our experiment results demonstrate the competitiveness of TA-FPN on both the MS COCO and PASCAL VOC datasets. Furthermore, ablation experiments underscore the effectiveness of each of the modules in detecting small objects. We hope that our work will contribute to further advancements in small object detection. In future research, we will study more deeply how to refine the small object features further when densely arranged and validate the generalization ability of our method in various backbone architectures and other vision-related tasks. Meanwhile, the design of a lightweight model structure will also be explored for easy deployment at the edge end to meet the needs of more practical scenarios.

Acknowledgement: The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Sheng Xiang: Formulation, Methodology, Writing original draft. Junhao Ma: Validation, Algorithm, Writing. Qunli Shang, Xianbao Wang and Defu Chen: Revising. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data and materials are available at https://pan.baidu.com/s/1Avyk-qUxStSF9vITqmfmbw?pwd=k7un (accessed 10/05/2024).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Qin Z, Zhou S, Wang L, Duan J, Hua G, Tang W. Motiontrack: learning robust short-term and long-term motions for multi-object tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Vancouver, Canada: IEEE. p. 17939–48.
2. Zheng Y, Liu X, Xiao B, Cheng X, Wu Y, Chen S. Multi-task convolution operators with object detection for visual tracking. IEEE Trans Circuits Syst Video Technol. 2022;32:8204–16. doi:10.1109/TCSVT.2021.3071128.
3. Jiao D, Fei T. Pedestrian walking speed monitoring at street scale by an in-flight drone. PeerJ Comput Sci. 2023;9(3):e1226. doi:10.7717/peerj-cs.1226.

4. Liu T, Du S, Liang C, Zhang B, Feng R. A novel multi-sensor fusion based object detection and recognition algorithm for intelligent assisted driving. IEEE Access. 2021;9:81564–74. doi:10.1109/ACCESS.2021.3083503.
5. 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. 2016;39(6):1137–49.
6. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Washington, USA: IEEE. p. 7263–71.
7. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco:common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland: Springer.
8. Khan AH, Nawaz MS, Dengel A. Localized semantic feature mixers for efficient pedestrian detection in autonomous driving. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023; Vancouver, Canada: IEEE. p. 5476–85.
9. Mohammed K, Abdelhafid M, Kassmi K, Ismail N, Atmane I. Intelligent driver monitoring system: an internet of things-based system for tracking and identifying the driving behavior. Comput Stand Interf. 2023;84:103704. doi:10.1016/j.csi.2022.103704.
10. Yang C, Huang Z, Wang N. QueryDet: cascaded sparse query for accelerating high-resolution small object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, USA. p. 13658–67.
11. Min K, Lee G-H, Lee S-W. ACNet: mask-aware attention with dynamic context enhancement for robust acne detection. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2021; Melbourne, Australia. p. 2724–2729.
12. Mahaur B, Mishra KK. Small-object detection based on YOLOv5 in autonomous driving systems. Pattern Recognit Lett. 2023;168:115–22. doi:10.1016/j.patrec.2023.03.009.
13. Huang H, Tang X, Wen F, Jin X. Small object detection method with shallow feature fusion network for chip surface defect detection. Sci Rep. 2022;12(1):3914. doi:10.1038/s41598-022-07654-x.
14. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Washington, USA: IEEE. p. 2117–25.
15. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell. 2015;38(1):142–58.
16. Girshick R. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision; 2015; Santiago, Chile: IEEE. p. 1440–8.
17. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017; Venice, Italy: IEEE. p. 2961–9.
18. 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; 2018; Salt Lake City, USA: IEEE. p. 6154–62.
19. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018.
20. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Vancouver, Canada: IEEE. p. 7464–75.
21. Bai Y, Zhang Y, Ding M, Ghanem B. Sod-mtgan: small object detection via multi-task generative adversarial network. In: Proceedings of the European Conference on Computer Vision, 2018; Munich, Germany: IEEE; p. 206–21.
22. Noh J, Bae W, Lee W, Seo J, Kim G. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019; Seoul, Republic of Korea: IEEE. p. 9725–34.

23. Bosquet B, Cores D, Seidenari L, Brea VM, Mucientes M, Del Bimbo A. A full data augmentation pipeline for small object detection based on generative adversarial networks. Pattern Recognit. 2023;133:108998. doi:10.1016/j.patcog.2022.108998.
24. Deng C, Wang M, Liu L, Liu Y, Jiang Y. Extended feature pyramid network for small object detection. IEEE Trans Multimed. 2021;24:1968–79.
25. Min K, Lee GH, Lee SW. Attentional feature pyramid network for small object detection. Neural Netw. 2022;155:439–50. doi:10.1016/j.neunet.2022.08.029.
26. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, USA: IEEE. p. 8759–68.
27. 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/CVF Conference on Computer Vision and Pattern Recognition; 2019; Los Angeles, USA: IEEE. p. 821–30.
28. Tan M, Pang R, Le QV. Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; Seattle, USA: IEEE. p. 10781–90.
29. 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; 2016; Las Vegas, NV, USA: IEEE. p. 770–8.
30. Luo Y, Cao X, Zhang J, Guo J, Shen H, Wang T, et al. CE-FPN: enhancing channel information for object detection. Multimed Tools Appl. 2022;81(21):30685–704. doi:10.1007/s11042-022-11940-1.
31. Guo G, Chen P, Yu X, Han Z, Ye Q, Gao S. Save the tiny, save the all: hierarchical activation network for tiny object detection. IEEE Trans Circuits Syst Video Technol. 2024;34(1):221–34. doi:10.1109/TCSVT.2023.3284161.
32. Zhao G, Ge W, Yu Y. GraphFPN: graph feature pyramid network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021; Montreal, Canada: IEEE. p. 2763–72.
33. Dong Y, Jiang Z, Tao F, Fu Z. Multiple spatial residual network for object detection. Complex Intell Syst. 2023;9:1347–62. doi:10.1007/s40747-022-00859-7.
34. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands: Springer.
35. Zhang S, Wen L, Bian X, Lei Z, Li SZ. Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, USA: IEEE. p. 4203–12.
36. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017; Venice, Italy: IEEE. p. 2980–8.
37. 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; 2020; Republic of Korea: IEEE. p. 9627–36.
38. Guo C, Fan B, Zhang Q, Xiang S, Pan C. AugFPN: Improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; Seattle, USA: IEEE. p. 12595–604.