mathematics

MDPI

Article

Composite Backbone Small Object Detection Based on Context and Multi-Scale Information with Attention Mechanism

Xinhan Jing, Xuesong Liu and Baolin Liu *

School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; m202110637@xs.ustb.edu.cn (X.J.); d202110404@xs.ustb.edu.cn (X.L.)
* Correspondence: liubaolin@ustb.edu.cn

Abstract: Object detection has gained widespread application across various domains; nevertheless, small object detection still presents numerous challenges due to the inherent limitations of small objects, such as their limited resolution and susceptibility to interference from neighboring elements. To improve detection accuracy of small objects, this study presents a novel method that integrates context information, attention mechanism, and multi-scale information. First, to realize feature augmentation, a composite backbone network is employed which can jointly extract object features. On this basis, to efficiently incorporate context information and focus on key features, the composite dilated convolution and attention module (CDAM) is designed, consisting of a composite dilated convolution module (CDM) and convolutional block attention module (CBAM). Then, a feature elimination module (FEM) is introduced to reduce the feature proportion of medium and large objects on feature layers; the impact of neighboring objects on small object detection can thereby be mitigated. Experiments conducted on MS COCO validate the superior performance of the method compared with baseline detectors, while it yields an average enhancement of 0.8% in overall detection accuracy, with a notable enhancement of 2.7% in small object detection.

Keywords: small object detection; context information; composite backbone network; multi-scale information; attention mechanism

MSC: 68T07

check for updates

Citation: Jing, X.; Liu, X.; Liu, B. Composite Backbone Small Object Detection Based on Context and Multi-Scale Information with Attention Mechanism. *Mathematics* **2024**, *12*, 622. https://doi.org/10.3390/math12050622

Academic Editors: Zaixing He, Mingqiang Wei, Qiong Wang and Meng Wu

Received: 21 January 2024
Revised: 14 February 2024
Accepted: 18 February 2024
Published: 20 February 2024

1. Introduction

Object detection is one fundamental field of computer vision, encompassing the identification of object positions, sizes, and categories within images. The precise detection of small objects, a specialized subtask within this field, is particularly crucial in diverse domains, playing a pivotal role in areas including autonomous driving, healthcare, and national defense. Its application in real-world scenarios is widespread and essential [1].

However, small object detection presents persistent challenges, arising from several factors. Firstly, small objects with limited resolution may result in restricted visual information, making it difficult for the detector to extract sufficient features. Furthermore, as small objects occupy a relatively tiny region, there is a lack of sufficient context information to assist in their detection. Additionally, the current detectors heavily depend on anchor boxes and use a fixed threshold during training to classify proposal regions as positive or negative. Therefore, an uneven distribution of positive and negative cases across various sizes will occur, resulting in fewer positive samples for small objects compared to medium and large ones. Consequently, detectors prioritize detecting other objects, often overlooking small ones [2,3].

Currently, research to facilitate small object detection primarily focuses on several areas: data augmentation, incorporating context information, and utilizing multi-scale information.

- Data augmentation [4–7] has been widely adopted in small object detection. Through utilizing diverse strategies to augment the training data, the dataset can be expanded

in size and diversity. However, data augmentation also presents certain issues, such as increased computational costs. And poorly devised augmenting approaches might add additional noise, negatively impacting feature extracting. Additionally, data augmentation merely adds features of small objects without considering how to optimize the extraction of these features.

- Certain studies have proposed the integration of context information to assist in detection tasks. This involves learning background features surrounding the object and global scene features. Though these explorations have yielded performance improvements, devising an appropriately balanced strategy for extracting context information and preventing small objects from being influenced by medium to large-sized objects remains a challenge.
- Moreover, multi-scale learning is widely used. The feature pyramid network (FPN) emerges as the multi-scale network for comprehensive feature extraction in object detection [8]. This approach aims to leverage an extensive range of feature layers, fusing the shallow layers and deep layer; the fused feature layer has richer position information and semantic information. Building upon this foundation, Liang et al. [9] proposed a Deep FPN, which incorporates lateral connections and is trained using specifically designed anchor boxes and loss functions. Merugu et al. [10,11] also employed a similar multi-module approach. However, these methods primarily focus on how to superimpose additional features for detection, ignoring specific multi-scale learning strategies tailored for small objects.

Although the current works are meaningful, there is still a lack of exploration regarding how to fully extract features and efficiently incorporate context information and multi-scale information. In light of these considerations, our study introduces a novel detection model for small objects. Initially, the model incorporates a composite backbone network, which can more thoroughly extract object features compared to current single-backbone network detectors. Additionally, the model is designed with a composite dilated convolution module, which efficiently integrates context information through dilated convolutions. Compared to existing context learning methods, this approach demonstrates a superior level of simplicity and efficiency. Furthermore, a feature reduction module is devised. Unlike existing multi-scale learning methods such as feature pyramids, this module effectively mitigates the impact of other-sized objects on small objects. This work conducts detailed experiments utilizing MS COCO, validating that our model is more efficient when detecting small objects than other comparative models.

The main contributions are listed below.

- This work introduces a composite backbone network architecture, enabling the two backbone networks to simultaneously extract and fuse features, thereby obtaining more usable features to enhance detection accuracy.
- This work designs a composite dilated convolution and attention module (CDAM). This module convolves and fuses shallow feature maps with varying dilation rates to effectively incorporate context information for better detection performance.
- This work presents the feature elimination module (FEM). This module mainly reduces the impact of medium and large objects on small objects by performing object elimination on the shallow feature layer.

2. Related Work

This part introduces the development of object detection. Subsequently, this paper delves into the detecting methods employed specifically towards small objects. Finally, the paper explores various approaches that utilize context information.

2.1. Object Detection

Detection models including R-CNN series [12] and the YOLO series [13–15] have shown commendable performance in precisely locating objects within images. Two-stage detectors, exemplified by the R-CNN series, typically exhibit superior detection accuracy.

They employ a two-step process, first selecting candidate boxes on feature layers and subsequently performing detection on these candidates. However, these detectors suffer from slow detection speeds, making them unsuitable for real-time applications. Conversely, one-stage detectors do not rely on generating candidate boxes on the feature map, resulting in faster detection speed; as a result, their accuracy typically falls short of that achieved by the former. Among one-stage detectors, YOLOv5 [16] has garnered significant attention. It adopts Mosaic data augmentation at the input, utilizes the Darknet-53 [17] architecture as the backbone, incorporates FPN and PANet [18] in the Head section, and introduces the GIOU_Loss function. Building upon YOLOv5, YOLOv7 enhances detection speed and accuracy through faster convolution operations and a more compact model with the same computational resources, surpassing many two-stage detectors. Despite the progress made by these detectors, small object detection remains a persistent challenge.

2.2. Small Object Detection

Recently, numerous methods have emerged to tackle the complexities associated with detecting small objects. These approaches encompass different aspects and considerations in order to enhance detection accuracy and overcome the limitations associated with small objects. A particular strategy involves utilizing data augmentation techniques to augment the representation of small objects and alleviate the impact of imbalanced sample distribution. Strategies such as copying, scaling, and component stitching have been explored for this purpose [4–6]. With the increasing popularity of reinforcement learning, research has gone beyond designing data augmentation strategies based solely on the features of small objects, and has instead explored using reinforcement learning to select the optimal data augmentation strategy [19], thus surpassing the limitations of object features. However, relying solely on data-level enhancements has limitations in significantly improving small object detection accuracy and may introduce unwanted noise. Another approach, known as multi-scale learning, combines spatial details in the shallow layers and semantic details in the deep layers to tackle the imbalance problem at the feature level [20–22]. The objective of this approach is to maximize the utilization of diverse scales, harnessing their complementary nature, with the ultimate goal of enhancing the detection ability of small objects. Additionally, contextual learning is another idea that has been explored for small object detection [23–26], which enriches object features by incorporating global or local context information explicitly or implicitly. By considering contextual cues, this approach enhances the discriminative power of features for small object detection. An additional approach involves utilizing generative adversarial networks (GANs) to generate high-resolution features based on low-resolution features as input, enriching small object features. However, due to the introduction of a more complex generative adversarial network, it has limitations of increased model complexity and slower speed. In addition, research has been conducted to overcome the limitations of anchor boxes. The existing design of anchor boxes makes the model more inclined to detect other objects and is not very helpful for small objects. Some studies have transformed the detection task into the estimation of keypoints. Law et al. [27] proposed CornerNet, which first predicts the two points on the diagonal and subsequently uses the paired corner points for the generation of the bounding box. Duan et al. [28] proposed CenterNet, which firstly predicts the two corner points as well as the center keypoint. Then, it matches the corners to form bounding boxes and finally utilizes the predicted center point to eliminate incorrect bounding boxes caused by corner mismatch. Yang et al. [29] introduced a novel technique called RepPoints, which offers a refined representation that enables more precise delineation of objects.

2.3. Context Information

For small object detection, incorporating context information holds significant importance. The limited resolution and pixel representation of small objects pose challenges for traditional object detection methods, impeding their accurate detection. Context information offers valuable cues that aid in this process. For instance, identifying a fish at the

bottom of the sea in an image might be arduous without any contextual cues. However, with the inclusion of ocean background information, the fish can be easily recognized. Therefore, integrating context information into the detection process proves advantageous for detecting small objects. Context information can be categorized as implicit or explicit. Implicit context encompasses the background features surrounding the object or global scene features. Li et al. [23] proposed a method that utilizes contextual windows; it extracts features through windows of different scales to introduce context information. Zeng et al. [24] tried to use Long Short-Term Memory between relevant image regions to extract context information. Explicit contextual reasoning entails leveraging clearly defined context information within a scene to aid in inferring the position or category of an object. For instance, the contextual relationship between the sky area and the object can be used to infer the object's category. Liu et al. [25] introduced a structural reasoning network that comprehensively considers the relationship between scene context and objects, thereby enhancing detection performance. To harness prior knowledge, Xu et al. [26] introduced Reasoning-RCNN, which is improved from Faster R-CNN. This model constructs a knowledge graph to encode contextual relationships and utilizes prior contextual relationships to influence the object detection process. Merugu et al. [30] integrated image classification with spatial context information by combining colorimetric edge preservation with spatial-spectral modeling.

The following mainly compares various object detection methods along three dimensions: feature extraction techniques, integration of context information, and utilization of multi-scale information, as illustrated in Table 1 below.

Table 1. Comparison of various approaches.

Method	Feature Extraction	Context Information	Multi-Scale Information
[12,31]	single backbone	-	-
[8]	single backbone	-	FPN
[15,16,18]	single backbone	-	based on FPN
[32]	single backbone	-	Img Pyramid
[23]	single backbone	by multi-context windows	-
[24]	single backbone	by LSTM	-
[25]	single backbone	by Graph Neural Network	-
[26]	single backbone	by Knowledge Graph	-
Ours	composite backbone	by dilated convolution	based on FPN and add FEM

3. Methodology

3.1. Overall Framework

The primary objective aims at improving detection ability for small objects by integrating context information, extracting more comprehensive features, and suppressing the impact of medium and large objects. To accomplish this, this work presents a novel small object detection model that integrates several crucial components. Firstly, the CDAM is introduced, which facilitates the comprehensive capture of context information. The module utilizes surrounding context and enhances the detection process. Secondly, the FEM is proposed, specifically designed to suppress the impact of medium and large object features. By focusing on small object detection, the FEM effectively eliminates the interference caused by larger objects and better detects small targets. Lastly, composite backbone networks are incorporated into our framework to enable more extensive and robust feature extraction. This integration of diverse backbone networks ensures the comprehensive representation of various object characteristics and enhances the overall detection performance. The architecture of the proposed method is illustrated in Figure 1.

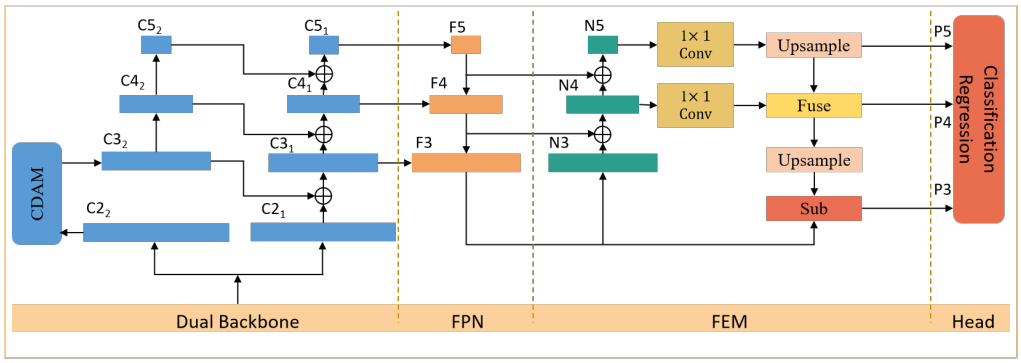

Figure 1. The architecture of the proposed method. Firstly, a composite backbone network is designed for extracting features. Backbone1 uses a basic backbone network, while Backbone2 adds the CDAM after the $C2_2$ feature layer in order to introduce context information; the rest of the network is the same as Backbone1. The composite backbone outputs three layers, F3, F4, and F5; the FEM is used to highlight features of small objects. Finally, P3, P4, and P5 are sent for further classification and regression.

3.2. Composite Dilated Convolution and Attention Module

Rich context information is helpful for small object detection. If the detector wants to obtain more context information, it needs a larger receptive field, which means the range of input space corresponding to a pixel on the output feature map.

By introducing gaps in the convolutional kernel, dilated convolution could achieve a better receptive field. This involves setting the dilation rate, specifying the gap between kernel elements during convolution. Dilated convolution has the advantage of preserving internal data structure and avoiding operations such as downsampling. Most importantly, the dilation rate governs the effective receptive field size of the convolutional operation. This empowers the model to gather information from a more extensive spatial context, enhancing its ability to capture surrounding details of objects in that region, referred to as context information—resulting in improved detection accuracy. However, a drawback of this approach arises from the discontinuity of the convolution kernel, leading to the exclusion of certain pixels during the calculation process. Consequently, there is a possibility of losing valuable pixel information.

In order to enrich context information without ignoring key details, this work proposes the CDAM that consists of two components: CDM and CBAM. Figure 2 illustrates how CDAM works.

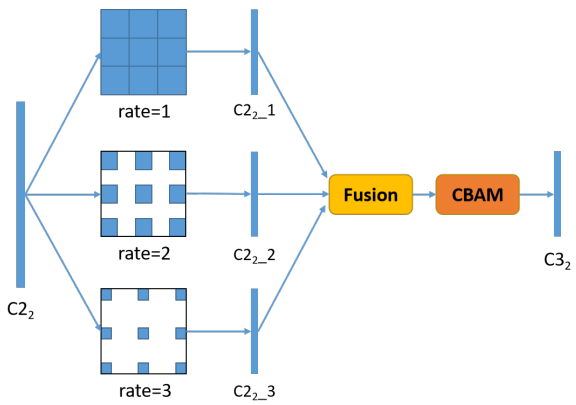

Figure 2. The composite dilated convolution and attention module (CDAM). Firstly, the $C2_2$ layer undergoes dilated convolution when three different rates of 1, 2, and 3 are set to obtain $C2_2$_1, $C2_2$_2, and $C2_2$_3. These three feature layers contain context information at different scales and are fused. To eliminate potential noise, the fused map is processed by the CBAM, obtaining $C3_2$.

The CDM contains composite dilated convolutions with different dilation rates, and then the composite feature map can be achieved through feature fusion. The CDM not only includes a wider range of context information but also integrates feature maps with diverse dilation rates, thus effectively avoiding the problem of critical information loss. This model adds the CDM after the feature layer $C2_2$. In the CDM, this model conducts dilated convolutions on the feature layer with a 3×3 convolution kernel. $C2_2$_1, $C2_2$_2, and $C2_2$_3 are separately obtained when the dilation rate sets to 1, 2, and 3, and are subsequently fused together. The fusion of the three feature layers is achieved by directly adding feature matrices, where addition is performed element-wise at each position. This approach allows preservation of the original range and distribution of the features.

The features processed by the CDM have richer context information, but at the same time, the fusion of multiple feature layers may introduce additional noise. Therefore, this model adds the CBAM after the CDM to prioritize effective information while further enriching the feature.

3.3. Feature Elimination Module

As the depth of the deep neural network increases, there is a gradual reduction in image resolution, resulting in the progressive loss of spatial information. Concurrently, the convolution operation may lead to the disappearance of many small objects. In contrast, medium and large objects retain more abundant features as semantic information gradually increases. Consequently, deep feature maps prove capable of meeting the detection requirements for medium and large objects.

However, shallow layers possess higher resolution, containing abundant spatial information and a substantial amount of object details, particularly pertaining to small objects. To highlight features of small objects, the FEM is introduced, shown in Figure 3. The FEM mainly reduces the proportion of medium and large objects in shallow feature maps that contain various object features. Through this, the proportion of positive samples for small objects can be increased, which alleviates the detection issues caused by sample imbalance. Combining the FEM, our approach directs the attention of the shallow layers towards detecting small objects. Consequently, this model achieves the goal of emphasizing detection of small objects without compromising detection of larger objects.

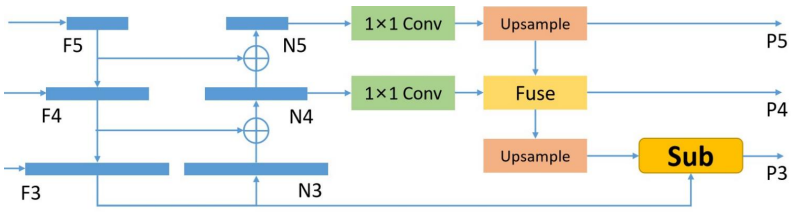

Figure 3. The feature elimination module (FEM). N4 and N5 are convolved with a 1×1 kernel to match the channel dimension of N3, then upsampled and fused to obtain the feature map N3′. Subtracting N3′ from N3 yields the feature map P3.

The FEM has three feature layers, F3, F4, and F5. As the convolutional layers go deeper, there is a reduction in the representation of features related to small objects, while the features associated with medium and large objects become more prominent. Therefore, in the shallow layer (F3), a feature layer containing rich small object features can be obtained. By utilizing deep features (F4, F5), the detector can effectively capture the features of other objects. After undergoing processing by the backbone network, they first go through a structure similar to PANet [18] to obtain the N3, N4, and N5 feature layers.

N3 is the shallow layer encompassing objects of all sizes. In order to retain the features of small objects, it is essential to suppress the features of objects of other sizes. By processing N4 and N5 according to the following formula, this model can obtain the feature layer that suppresses small object features and highlights medium and large object features, denoted as N3′.

$$N3' = U[U(Conv(N5)) \oplus Conv(N4)] \quad (1)$$

In this formula, U denotes upsampling, a technique employed to increase resolution of an image. This process entails the insertion of additional data points between existing ones, thereby expanding the overall size of the image. Specifically, nearest-neighbor interpolation is employed, which can leave the pixel values of the original image unchanged while directly replicating the pixel values of the nearest neighbors to the new pixel positions. Consequently, this approach effectively conserves the content of the image, thereby maintaining the object shapes and edge details without introducing blurring effects. This is helpful for the recovery of medium to large-sized objects. N3′ almost only contains medium and large objects, with the large-sized objects being similar to those in N3. To derive a feature layer exclusively containing small objects, the following adjustment is made.

$$\mathrm{P3} = \mathrm{N3} \ominus \mathrm{N3}' \tag{2}$$

P3 suppresses the features of objects of other sizes, exclusively preserving the characteristics of small ones. This adjustment introduces a bias towards small objects, thereby enhancing the model's capability to detect small objects.

3.4. Composite Backbone Networks

In contrast to objects of other sizes, small objects occupy a relatively smaller proportion of pixels, making their features difficult to extract, which becomes a key limiting factor for small object detection accuracy. Currently, most detectors are based on single backbone networks. Although single backbone networks have lower computational complexity and relatively higher efficiency, their feature extraction capabilities are limited. They cannot fully utilize the advantages of different feature extraction methods, making it difficult to achieve diversified feature fusion, which limits the representation and detection performance. In contrast, composite backbone networks, by integrating two different feature extraction networks, can provide richer and more diverse feature representations, thereby enhancing the capability of representing features. Moreover, the interaction between different backbones can effectively improve the robustness of the model. If one backbone performs poorly under certain circumstances, the other backbone will provide an alternative feature representation. These two sets of features complement and strengthen each other, enhancing the stability of the model.

To solve this, this study introduces a composite backbone network with specific designs. By jointly extracting object features from images, the model's representation capability can be enhanced, which could better capture small object features. Figure 4 shows the structure of the specific network, consisting of Backbone1 and Backbone2. The two backbone networks have different structures, with Backbone1 being the base backbone network following a traditional structure. As for Backbone2, the CDAM is added after the $C2_2$ feature layer, while the other parts are the same as Backbone1. The two distinct backbone networks maintain separate weight parameters. However, Backbone1 integrates the feature layers extracted from Backbone2 and leverages them in the training process for subsequent feature layers. This non-sharing of weights allows each backbone network to learn features independently, thereby enabling the model to concentrate on diverse levels and facets of feature extraction. This approach enriches the representation of features and reduces the risk of overfitting. The composite backbone networks are processed according to the following formula.

$$\mathrm{C3}_2 = \mathrm{CDAM}(\mathrm{C2}_2) \tag{3}$$

$$\mathrm{C3}_1 = \mathrm{Conv}(\mathrm{C2}_1) \oplus \mathrm{C3}_2 \tag{4}$$

$$\mathrm{C4}_1 = \mathrm{Conv}(\mathrm{C3}_1) \oplus \mathrm{Conv}(\mathrm{C3}_2) \tag{5}$$

$$\mathrm{C5}_1 = \mathrm{Conv}(\mathrm{C4}_1) \oplus \mathrm{Conv}(\mathrm{Conv}(\mathrm{C3}_2) \tag{6}$$

Firstly, $C2_2$ is processed by the CDAM to obtain feature layer $C3_2$; then, $C3_2$ and convolved $C2_1$ are fused to obtain feature layer $C3_1$. Both $C4_2$ and $C5_2$ are obtained by convolution on their previous layers. The acquisition process of $C4_1$ and $C5_1$ is similar to $C3_1$, obtained by

fusing $C4_2$ and $C5_2$, respectively, with the convolved $C3_1$ and $C4_1$. This interactive feature extraction process allows the two backbone networks to complement each other's features to the fullest extent. Finally, the composite backbone network outputs three feature layers, F3, F4, and F5, as the input of the FPN section shown in Figure 1.

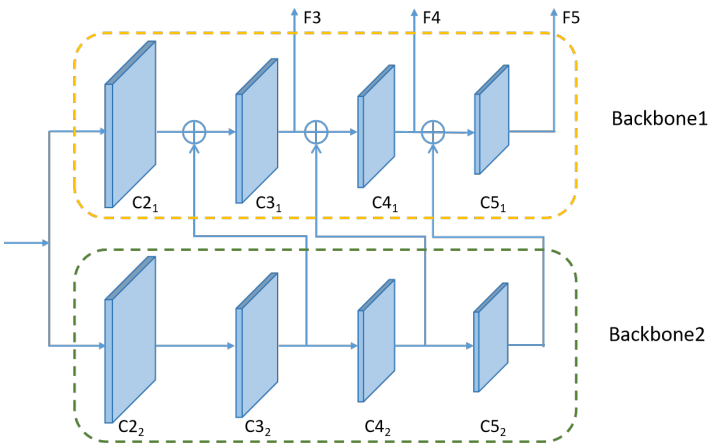

Figure 4. The composite backbone module. Backbone1 uses a basic backbone network; Backbone2 adds the CDAM after the $C2_2$ feature layer, as indicated by the yellow box in the figure. The two backbone networks are separately trained without sharing parameters, and the feature layers of each network are fused and used for training the next layers. Finally, three feature maps, F3, F4, and F5, are output.

Based on the above composite backbone structure, the two backbone networks extract features in different ways and then fuse them, enriching the object feature information and improving small object detection accuracy. Figure 5 describes the entire procedure.

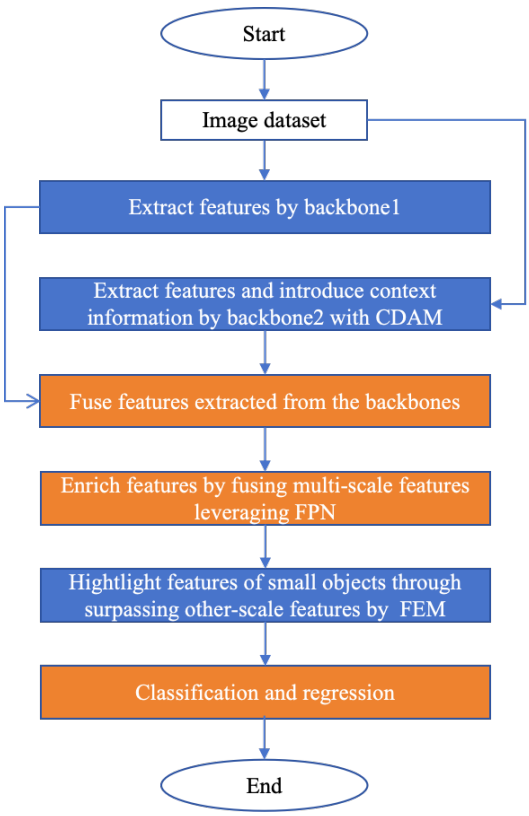

Figure 5. Flowchart of the proposed method.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

The experiments are conducted on MS COCO (Microsoft Common Objects in Context) [33], which serves as a benchmark for object detection. MS COCO is a comprehensive image dataset encompassing a wide variety of objects. It comprises images featuring 80 distinct object categories, containing training, validation, and test sets (with more than 118,000, 5000, and 40,000 images, respectively).

First, the training process is executed utilizing the training set. Then, the ablation experiments are carried out specifically using the validation set to verify the performance of our proposed detection model. Finally, the model undergoes testing on the test set, where it is compared against other models for performance evaluation.

4.1.2. Performance Indicator

Average precision (AP) is a primary evaluation indicator for MS COCO. AP is computed by taking the average of precision values at different Intersection over Union (IoU) thresholds. These thresholds are typically in the range of [0.5, 0.95], with an increment of 0.05. In MS COCO, around 41% of the objects fall into the category of small objects (with an area less than 32^2), 34% are classified as medium objects (with an area between 32^2 and 96^2), and approximately 24% are considered large objects (with an area exceeding 96^2). On this basis, different measurement standards for these objects are proposed, including AP_S, AP_M, and AP_L, indicating average precision for small, medium, and large objects. In the experiments, AP_S is used to evaluate the proficiency of the model in detecting small objects. Simultaneously, this work observes AP_M and AP_L to demonstrate that the model maintains good performance in medium and large object detection.

4.1.3. Training Details

All of our experiments are performed utilizing three NVIDIA 2080Ti GPUs. To optimize the model, the Adam optimizer is employed with default parameters provided by PyTorch 1.7.0. The batch size is set as 4 during training. Initially, the learning rate is 0.001, while at epoch 10, epoch 20, and epoch 30, the rate is divided by 0.1. This work maintains consistency with YOLOv7 by keeping all other parameters unchanged. The model undergoes training for 40 epochs. Figure 6 depicts the loss curve, which demonstrates that our model has good convergence.

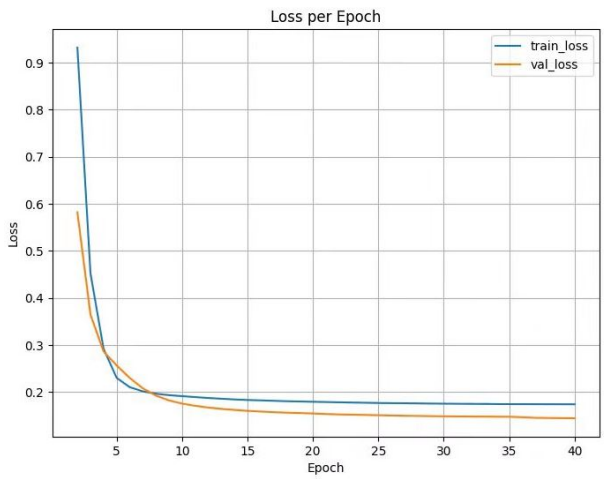

Figure 6. The loss across different epochs for our model.

For ablation experiments, the input size of the images is set as 416 × 416 pixels.

4.2. Ablation Study

4.2.1. The Impact of CDAM

As aforementioned, object detection can benefit from additional context information; thus, the CDAM is proposed as an efficient approach to incorporate maximum context information. After incorporating the composite dilated convolution, this study further introduces CBAM to mitigate the impact of introduced noise and enhance the model's focus on crucial regions.

First, this work conducts some ablation studies in order to validate the efficiency of the CDAM. Table 2 shows the outcome. Compared to the baseline, the inclusion of the composite dilated convolution shows an improvement of 0.6% in detecting accuracy, while incorporating the CBAM results in an additional improvement of 0.3%. Moreover, with the introduction of the CDAM, while AP_S increase, there are corresponding improvements in AP_M and AP_L, rising by 2.2% and 3%, respectively. The above results conclusively demonstrate the efficacy of the CDAM, as well as the effectiveness of the composite dilated convolution. These enhancements not only enable the model to leverage context information more effectively but also strengthen its precise focus on crucial regions, providing clear evidence for the improvement in object detection performance across various object sizes.

Table 2. CDAM ablation study on MS COCO val2017 Dataset. Baseline: YOLOv7 trained by ourselves. The CDM: only using the composite dilated convolution operation. CDM + CBAM: adding the CBAM on the basis of the CDM.

Method	Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	416^2	43.6	62.2	47.0	22.9	45.2	59.9
CDM	416^2	44.4	62.6	48.3	23.5	46.6	60.7
CDM + CBAM	416^2	44.9	63.0	48.8	23.8	47.4	62.9

Additionally, this study conducts experiments on combinations of different dilation rates to assess our method. According to observations from Table 3, it is evident that when rate = 1, the detection accuracy is the lowest. However, upon combining different dilation rates, there is an observable enhancement in accuracy. The highest detection accuracy is achieved when all three dilation rates are combined. This is attributed to the diverse context information introduced by the three different dilation rates, thus enhancing the detection accuracy across different sizes of objects.

Table 3. Ablation study on combinations of different dilation rates in the CDM.

Rate = 1	Rate = 2	Rate = 3	Size	AP_S	AP_M	AP_L
✓	-	-	416^2	22.9	45.2	59.9
✓	✓	-	416^2	23.3	45.9	60.1
✓	-	✓	416^2	23.2	45.7	60.5
✓	✓	✓	416^2	23.5	46.6	60.7

Tables 2 and 3 demonstrate the effectiveness of the CDAM. The efficacy is attributed to the preservation of contextually relevant information surrounding small objects within feature layer $C2_2$, which remains unchanged by convolution operations. This preservation facilitates the detection of small objects by maintaining beneficial context cues. Furthermore, employing dilated convolutions on $C2_2$ enhances the receptive field of the subsequent feature layer $C3_2$. Consequently, $C3_2$ can effectively capture features of small objects from the preceding layer while concurrently considering context information pertinent to these objects. The incorporation of such information serves as a mechanism for feature enhancement, thereby facilitating the detection of small objects. Moreover, the utilization of two distinct dilation rates (2 and 3) equates to the establishment of two disparate

sizes of information extraction windows. This approach effectively enriches the pool of contextual information, thereby augmenting the overall efficacy of the framework in small object detection.

4.2.2. The Impact of FEM

Within the shallow feature layer, there exists abundant information about the features of objects. However, for small objects, the presence of extensive features from medium and large objects often diverts the model's attention away from the small objects. This work hypothesizes that by eliminating the medium and large object features within the shallow feature layer, the model can prioritize and concentrate on detecting small objects more effectively.

Upon the introduction of the feature elimination module, the detection precision of small objects notably increases from 22.9% to 23.7%, as depicted in Table 4. Meanwhile, large object detection exhibits some degree of enhancement as well. This is because in the deep feature map, large object features are already distinctive enough to support the effective detection of large objects. For medium-sized objects, the detection precision remains roughly stable. The results show that the FEM could facilitate small object detection without negative impact on detecting other-sized objects.

Table 4. FEM ablation study on MS COCO val2017 dataset.

Method	Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	416^2	43.6	62.2	47.0	22.9	45.2	59.9
FEM	416^2	44.7	63.0	48.5	23.7	45.1	60.3

The effectiveness of the FEM in detecting small objects is attributed to the fact that the F3 feature layer undergoes fewer convolution operations compared to F4 and F5. This results in diminished loss of features and richer object information, especially small object information. Through feature elimination, the method selectively suppresses information concerning objects of other sizes within the F3 feature layer. Therefore, after training, the model weights within this layer are more favorable for detecting small objects. Meanwhile, despite F4 and F5 undergoing convolution operations wherein certain features of small objects may be lost, features of medium and large objects still dominate; hence, the detection of medium and large objects is not affected.

4.2.3. The Impact of Composite Backbone

To better extract object features, this model incorporates a composite backbone network structure. Table 5 shows the results.

Table 5. Composite backbone ablation study on MS COCO val2017 dataset.

Method	Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	416^2	43.6	62.2	47.0	22.9	45.2	59.9
Composite Backbone	416^2	44.7	63.1	48.4	24.0	46.2	61.3

Upon the incorporation of the composite backbone structure, for small objects, the precision increases from 22.9% to 24%, resulting in an improvement of 1.1%. Additionally, the precision for medium objects increases by 1%, while that for large objects increases by 1.4%. These experimental findings clearly demonstrate the substantial benefits of the composite backbone network in feature extraction, validating its effectiveness in enhancing overall detection performance.

The composite backbone contributes to the improvement of detection accuracy for several reasons. Firstly, the utilization of two different backbone networks enables the provision of more diverse and enriched feature representations, thereby enhancing the

model's capability to represent features. Additionally, the composite backbone network can enhance the robustness of the model by facilitating interactive learning between different backbone structures. If one backbone structure performs poorly under certain conditions, the other backbone structure can provide better feature representations, thereby enhancing the stability of the model.

However, the introduction of composite backbone networks leads to an increase in computational complexity, resulting in a certain degree of reduction in training speed. Under the same training conditions, the computational complexity of composite backbone networks is 1.63 times that of single backbone networks, leading to a final training speed that is 1.43 times slower compared to single backbone networks.

4.3. Comparison with Baseline

In this study, the model selects YOLOv7 as the baseline model and implements several improvements upon it. To ensure a fair and effective comparison with YOLOv7, this work utilizes identical parameters and configurations for both models. The training and evaluation of YOLOv7 and our model are performed using three NVIDIA 2080Ti GPUs, with a batch size of 4. The Adam optimization algorithm was employed during training, and the MS COCO train2017 dataset was used. Experiments were conducted on the MS COCO val2017 dataset. The comparison results between YOLOv7 and our model are presented in Table 6.

Table 6. Comparison between YOLOv7 and our model on MS COCO val2017 dataset.

Method	Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
YOLOv7	640^2	45.8	65.5	49.2	27.1	52.3	62.5	54.7
Ours	640^2	46.5	65.2	51.1	29.7	51.9	62.6	55.6

Compared with the baseline, our method demonstrates a notable enhancement when detecting small objects, with a 2.6% increase in AP_S. In the detection of large objects, our method performs roughly on par with the baseline. This evidence confirms that our approach provides a notable enhancement in small object detection. The overall improvements in AP and AR also demonstrate the effectiveness of the proposed method. Figure 7 shows precision-recall curves and F1 curves.

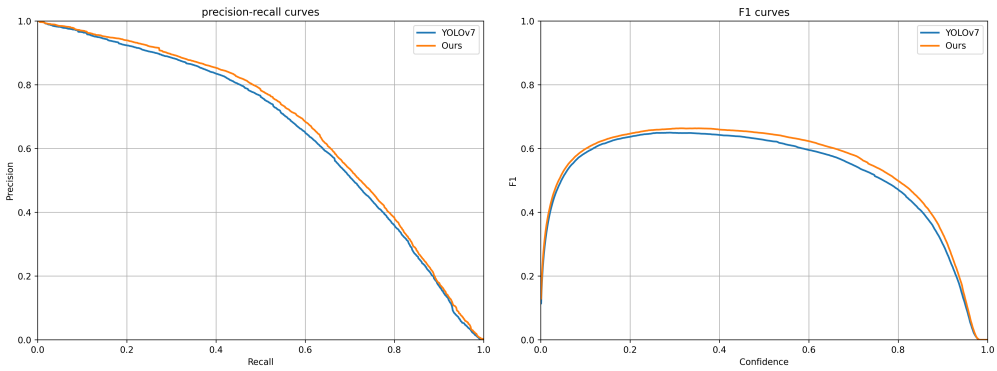

Figure 7. Precision-recall curves and F1 curves of our model and YOLOv7.

4.4. Comparison with Other Models

Finally, the proposed methodology is evaluated against related approaches on the MS COCO test set, with results presented in Table 7. The comparative experiments categorize the input image sizes into two groups: 640 × 640 and 512 × 512. When image size is set as 640 × 640, our model achieves an AP_S of 27.2% and an AP of 47.2%. Conversely, with an input image size of 512 × 512, the AP_S reaches 24.1% and the AP is 46.1%. These findings highlight the effectiveness of our model across various image sizes.

Table 7. Comparison of results between our method and related methods on MS COCO test2017 dataset.

Method	Backbone	Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [34]	ResNet50	700^2	35.1	54.2	37.7	18.0	39.3	46.4
Faster R-CNN + FPN [8]	ResNet101	-	36.2	59.1	39.0	18.2	39.0	48.2
Cascade R-CNN [35]	ResNet101	-	42.8	62.1	46.3	23.7	45.5	55.2
Soft-NMS [36]	Aligned-Inception-ResNet	800^2	40.9	62.8	-	23.3	43.6	53.3
Grid R-CNN + FPN [37]	ResNeXt101	800^2	43.2	63.0	46.6	25.1	46.5	55.2
LH R-CNN [38]	ResNet101	800^2	41.5	-	-	25.2	45.3	53.1
IPG R-CNN [32]	IPGNet101	800^2	45.7	64.3	49.9	26.6	48.6	58.3
Ours	CSPDarknet53	640^2	47.2	65.5	51.8	27.2	51.8	61.1
YOLOv2 [13]	Darknet	544^2	21.6	44.0	19.2	5.0	22.4	35.5
SSD [31]	ResNet101	512^2	31.2	50.4	33.3	10.2	34.5	49.8
DSSD [39]	ResNet101	512^2	33.2	53.3	35.2	13.0	35.4	51.1
DES [40]	VGG16	512^2	30.1	53.2	34.6	13.9	36.0	47.6
DFPR [41]	ResNet101	512^2	34.6	54.3	37.3	14.7	38.1	51.9
RefineDet [42]	VGG16	512^2	33.0	54.5	35.5	16.3	36.3	44.3
CenterNet [28]	HRNet-W64	512^2	44.0	62.6	47.1	23.0	47.3	57.8
CornerNet [27]	Hourglass104	512^2	42.1	57.8	45.3	20.8	44.8	56.7
Ours	CSPDarknet53	512^2	46.1	64.7	50.5	24.1	50.9	62.8

In Table 8, the proposed method is compared with QueryDet [43], which is the SOTA method. QueryDet conducted experiments with both ResNet50 and ResNet101, while our method is trained with CSPDarknet53. In terms of small object detection accuracy, the proposed method outperforms QueryDet with ResNet50 by 1.8%. However, compared to QueryDet with ResNet101, the accuracy is slightly lower, as deeper networks often yield greater performance improvements, especially for hard-to-detect small objects. Nevertheless, our method performs better on objects of other sizes. On the other hand, the introduction of composite backbone networks brings additional computational costs and parameters, resulting in an FPS of 11.58, which is inferior to the 14.88 of QueryDet. Additionally, our method consists of three modules designed specifically for small objects, leading to a noticeable improvement in accuracy. But in terms of implementation convenience, there is still a gap between the proposed method and the SOTA method. Therefore, reducing computational complexity while maintaining accuracy and improving efficiency is the direction we aim to optimize in future work.

Table 8. Comparison with other SOTA methods on MS COCO test2017 dataset.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS	Params (M)
QueryDet [43]	ResNet50	41.6	62	44.5	25.4	43.8	51.2	14.88	37.74
QueryDet [43]	ResNet101	43.8	64.3	46.9	27.5	46.4	53	-	-
Ours	CSPDarknet53	47.2	65.5	51.8	27.2	51.8	61.1	11.58	49.59

4.5. Qualitative Results

The qualitative results, presented in Figure 8, provide a visual comparison between YOLOv7 and the proposed model. Notably, the proposed model exhibits the ability to detect a greater number of objects, while exhibiting a distinct advantage for detecting small objects in particular.

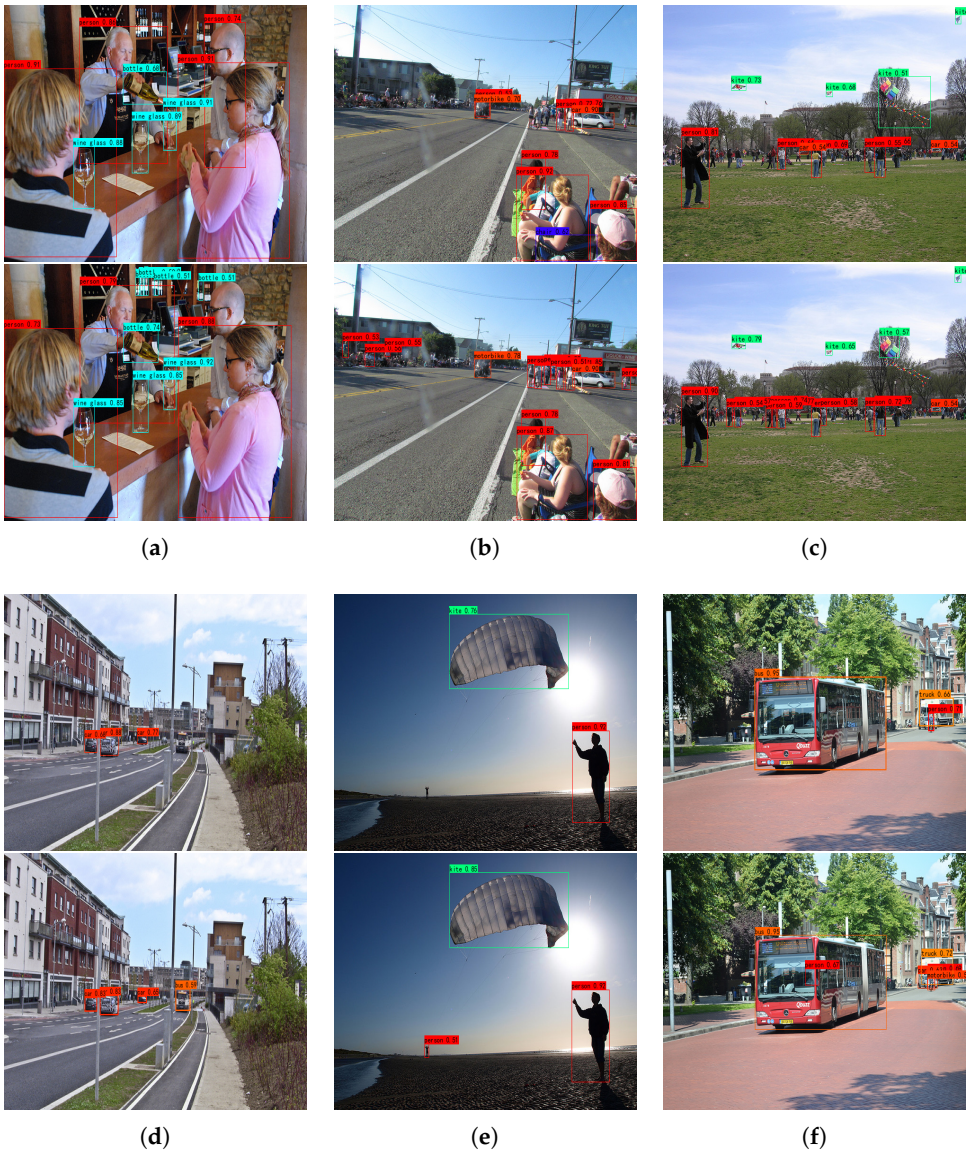

Figure 8. Qualitative results comparison between baseline and our method. The images in the top row are the results of the baseline, while the ones on the bottom are achieved by our method.

5. Conclusions and Future Works

This study proposes a method that utilizes a composite backbone network, context information, and multi-scale information. It employs the composite backbone network to extract richer and more informative features, enhancing the representation capability of our model. Additionally, the model introduces the CDAM, which efficiently incorporates context information through dilated convolutions at different ratios and reduces noise interference through the attention mechanism. Finally, this study designs the FEM to fully utilize multi-scale features, effectively mitigating the impact brought by medium to large-sized objects on small object detection. Experiments conducted on MS COCO demonstrate the roles of each module and the superiority of the overall model in small object detection.

However, it is important to note that the introduction of composite backbone networks, while enhancing detection accuracy, does lead to increased computational complexity, resulting in decreased training and inference speeds. Future research could explore strategies to alleviate this computational overhead, such as optimization backbone network architectures. Additionally, further investigation into the generalization ability of the proposed model across diverse datasets and object categories would be valuable. Moreover, exploring

the potential integration of context information techniques and enhancing the adaptability of the model to varying environmental conditions could contribute to its robustness and applicability in real-world scenarios.

Author Contributions: Conceptualization, X.J. and X.L.; methodology, X.J.; validation, X.J.; writing—original draft preparation, X.J.; writing—review and editing, X.J., X.L. and B.L.; supervision, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No.U2133218), the National Key Research and Development Program of China (No.2018YFB0204304), and the Fundamental Research Funds for the Central Universities of China (No.FRF-MP-19-007 and No.FRF-TP-20-065A1Z).

Data Availability Statement: All data used in this study are publicly available and can be accessed directly from MS COCO (Microsoft Common Objects in Context, https://cocodataset.org/).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
2. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards large-scale small object detection: Survey and benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]
3. Zhu, Y.; Zhou, Q.; Liu, N.; Xu, Z.; Ou, Z.; Mou, X.; Tang, J. ScaleKD: Distilling Scale-Aware Knowledge in Small Object Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19723–19733.
4. Kisantalk, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
5. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Wang, X.; Sun, X.; Dong, J. Rrnet: A hybrid detector for object detection in drone-captured images. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
6. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Xiang, S.; Sun, J.; Jia, J. Stitcher: Feedback-driven data provider for object detection. *arXiv* **2020**, arXiv:2004.12432.
7. Demirel, B.; Baran, O.B.; Cinbis, R.G. Meta-tuning Loss Functions and Data Augmentation for Few-shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7339–7349.
8. Lin, T.Y.; 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
9. Liang, Z.; Shao, J.; Zhang, D.; Gao, L. Small object detection using deep feature pyramid networks. In *Advances in Multimedia Information Processing–PCM 2018, Proceedings of the 19th Pacific-Rim Conference on Multimedia, Hefei, China, 21–22 September 2018*; Proceedings, Part III 19; Springer International Publishing: Cham, Switzerland, 2018; pp. 554–564.
10. Bathula, A.; Muhuri, S.; Gupta, S.K.; Merugu, S. Secure certificate sharing based on Blockchain framework for online education. *Multimed. Tools Appl.* **2023**, *82*, 16479–16500. [CrossRef]
11. Bathula, A.; Merugu, S.; Skandha, S.S. Academic Projects on Certification Management Using Blockchain—A Review. In Proceedings of the 2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC), Hyderabad, India, 28–30 December 2022; pp. 1–6.
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
13. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
14. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
15. Wang, C.Y.; Bochkovskiy, A.; Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
16. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. *arXiv* **2021**, arXiv:2108.11539.
17. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv*, **2018**, arXiv:1804.02767.
18. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

19. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.Y.; Shlens, J.; Le, Q.V. Learning data augmentation strategies for object detection. In *Computer Vision–ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXVII 16; Springer International Publishing: Cham, Switzerland, 2020; pp. 566–583.
20. Nayan, A.A.; Saha, J.; Mozumder, A.N.; Mahmud, K.R. Real time detection of small objects. *Int. J. Innov. Technol. Explor. Eng.* **2020**, *9*, 837–843.
21. Liu, Z.; Gao, G.; Sun, L.; Fang, Z. HRDNet: High-resolution detection network for small objects. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
22. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [CrossRef]
23. Li, J.; Wei, Y.; Liang, X.; Dong, J.; Xu, T.; Feng, J.; Yan, S. Attentive contexts for object detection. *IEEE Trans. Multimed.* **2016**, *19*, 944–954. [CrossRef]
24. Zeng, X.; Ouyang, W.; Yan, J.; Li, H.; Xiao, T.; Wang, K.; Liu, Y.; Zhou, Y.; Yang, B.; Wang, Z.; et al. Crafting gbd-net for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2109–2123. [CrossRef] [PubMed]
25. Liu, Y.; Wang, R.; Shan, S.; Chen, X. Structure inference net: Object detection using scene-level context and instance-level relationships. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6985–6994.
26. Xu, H.; Jiang, C.; Liang, X.; Lin, L.; Li, Z. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6419–6428.
27. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
28. 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, 27–28 October 2019; pp. 6569–6578.
29. 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, 27–28 October 2019; pp. 9657–9666.
30. Merugu, S.; Tiwari, A.; Sharma, S.K. Spatial–spectral image classification with edge preserving method. *J. Indian Soc. Remote Sens.* **2021**, *49*, 703–711. [CrossRef]
31. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
32. Liu, Z.; Gao, G.; Sun, L.; Fang, L. IPG-net: Image pyramid guidance network for small object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 1026–1027.
33. Available online: https://cocodataset.org/ (accessed on 1 January 2017).
34. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
35. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
36. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
37. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7363–7372.
38. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Light-head r-cnn: In defense of two-stage object detector. *arXiv* **2017**, arXiv:1711.07264.
39. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
40. Zhang, Z.; Qiao, S.; Xie, C.; Shen, W.; Wang, B.; Yuille, A.L. Single-shot object detection with enriched semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5813–5821.
41. Kong, T.; Sun, F.; Tan, C.; Liu, H.; Huang, W. Deep feature pyramid reconfiguration for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 169–185.
42. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
43. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13668–13677.