sustainability

MDPI

Article

Applying an Intelligent Approach to Environmental Sustainability Innovation in Complex Scenes

Hongjie Deng, Daji Ergu *, Fangyao Liu, Bo Ma and Ying Cai

Key Laboratory of Electronic and Information Engineering (Southwest Minzu University), State Ethnic Affairs Commission, Chengdu 610041, China
* Correspondence: ergudaji@163.com

Abstract: Environmental protection is still a key issue that cannot be ignored at this stage of social development. With the development of artificial intelligence, various technologies increasingly tend to be widely used in the field of environmental protection, such as searching the wilderness through an unmanned aerial vehicle (UAV) and cleaning garbage by robots. Traditional object detection algorithms for this scenario suffer from low accuracy and high computational cost. Therefore, this paper proposes an algorithm applied to automatic garbage detection and instance segmentation in complex scenes. First, we construct sample-fused feature pyramid networks (SF-FPN) to achieve multi-scale feature sampling on multiple levels, to enhance the semantic representation of features. Second, adding the mask branch based on conditional convolution, introducing the idea of instance-filters to automatically generate the filter parameters of the Fully Convolutional Networks (FCN), to realize the instance-level pixel classification. Moreover, the Atrous Spatial Pyramid Pooling (ASPP) module is introduced to encode the feature information in a dense way to assist the generation of MASK. Finally, the object is detected and the instance is segmented by a two-branch structure. In addition, we also perform data augmentation on the original dataset to prevent model overfitting. The proposed algorithm reaches 82.7 and 72.4 according to the mAP index of detection and instance segmentation while using the public TACO dataset.

Keywords: deep learning; object detection; instance segmentation; environmental sustainability

Citation: Deng, H.; Ergu, D.; Liu, F.; Ma, B.; Cai, Y. Applying an Intelligent Approach to Environmental Sustainability Innovation in Complex Scenes. *Sustainability* **2022**, *14*, 16758. https://doi.org/10.3390/su142416758

Academic Editor: Gwanggil Jeon

Received: 26 November 2022
Accepted: 12 December 2022
Published: 14 December 2022

1. Introduction

With the continuous realization of human activities, the earth has been extremely damaged, and if urgent and stronger actions are not taken to protect the environment, the earth's ecosystem and the cause of sustainable development of humans will be increasingly threatened. In the face of increasing garbage production, how to maximize the use of waste resources, and reduce the amount of garbage disposal to improve the quality of the living environment is currently one of the urgent issues of common concern in the world. Since 1 May 2020, Beijing has proposed to separate domestic waste. It shows that the task of garbage sorting is an integral part of social development at this stage [1–4].

With the development of deep learning, computer vision has become a popular field of artificial intelligence at this stage, in which technologies such as object detection and instance segmentation have also contributed to the development of daily life toward intelligence.

At this stage, object detection algorithms are divided into one-stage and two-stage algorithms.

The One-stage algorithm directly regresses the positions and the class probabilities of objects. Its representative detection algorithms have the Yolo family of algorithms (Yolov1 [5], Yolov2 [6], Yolov3 [7], Yolov4 [8], etc.) and SSD [9]. Yolov1 divides the whole graph into S × S grids and performs edge and category prediction for the targets in the grids. Finally, the optimal edges are obtained by NMS. Both Yolov2 and SSD use a priori frames, which cover different positions of the whole image by pre-setting a set of borders

with different sizes and aspect ratios for each grid. Yolov3 uses Darknet-53 as the backbone and introduces Feature Pyramid Networks (FPN) [10] which can better correspond to objects of different sizes for multi-scale prediction. The idea of Yolo4 is to find the best balance between the input network resolution, the number of convolutional layers, the number of parameters, and the number of layer outputs.

The two-stage algorithm divides the detection task into two phases, i.e., generating candidate regions (Region proposals) and classifying the candidate regions with border regression. Such algorithms are typically represented by the R-CNN family (R-CNN [11], Fast R-CNN [12], Faster R-CNN [13], Mask R-CNN [14], etc.). R-CNN generates about two thousand candidate frames on the original image with Selective Search [15], and finally, the target classification and edge regression are achieved by a single classification SVM. Fast R-CNN replaces the object for selective search with a feature map generated by convolution operation based on the former and uses RoI pooling to make the generated RoI mapping a fixed-size feature map. To improve the speed of detection box generation, Faster R-CNN designed the RPN module, which achieves regional proposal generation by setting multiple scales of an anchor box. In contrast, the instance segmentation task is to generate the mask corresponding to the object by pixel-level classification based on object detection.

For previous research of the R-CNN family of algorithms, Kaiming He proposed Mask R-CNN, which introduces the mask branch (composed of FCN [16]) based on Faster R-CNN, and performs pixel-by-pixel prediction of the object by a fully convolutional network, thus realizing the instance segmentation under object detection.

Therefore, in this paper, considering object detection in complex scenes and weighing the accuracy and speed of task implementation, we propose an anchor-free algorithm for garbage detection to achieve the tasks of object detection and instance segmentation simultaneously.

This paper contributes as follows:

- The sample-fused FPN is constructed to achieve feature extraction and fusion at multiple scales, enhancing the semantic information of the features.
- Adding a mask branch based on conditional convolution [17]. By ASPP [18], the module encodes the feature values in a dense form, assisting mask generation, and automatically generates filter parameters for different instances based on conditional convolution, thus the corresponding mask generation is achieved.
- The dual branches of mask and box are trained in an end-to-end form, based on which the object detection and instance segmentation tasks are implemented.

Combining the above improvements, the model in this paper is trained and tested on the public dataset TACO [19]. Compared with the mainstream algorithms at this stage, the model presents better results for the task and can achieve accurate garbage detection and instance segmentation.

2. Related Works

As a hot topic in urban governance at this stage, researchers have proposed relevant solutions based on computer vision. Lee used SSD with AlexNet as the backbone network for waste detection [20]. Ma proposed a lightweight Single-Stage Detector (SSD) with a novel feature fusion model for solving the problem of garbage detection [21]. Cao migrated the InceptionV3 model, which has been pre-trained on the ImageNet dataset, to the task of garbage recognition [22]. Kang used ResNet-34 network as the base architecture and improved the three aspects of feature fusion, residual unit, and the activation function to achieve garbage classification [23]. Shi improved the accuracy of garbage classification by branching extensions to the Xception and implementing multi-layer network feature fusion [24]. However, the above network models are limited to garbage detection in a simple context and are only trained and tested on single-object images without considering the numerous visual interference elements in natural scenes, e.g., the presence of occlusions, the size of garbage volume, etc. As artificial intelligence gradually replaces human work in many aspects, how to implement garbage detection in natural environments must become a key task.

Therefore, this paper improves the FCOS network structure based on its base module. We replace traditional FPN by building SF-FPN structure to enhance the capability of feature extraction. The idea of conditional convolution is introduced to automatically generate the network parameters and reduce the computational cost. Eventually, the two-branch structure of detection and instance segmentation is trained end-to-end so that the network can achieve the corresponding tasks simultaneously.

3. FCOS

FCOS is a one-stage algorithm proposed by Tian to solve the object detection problem on a pixel-by-pixel method [25]. The network structure is shown in Figure 1. The core idea of the algorithm is to directly regress the class and minimum bounding box of the object corresponding to the current position at each pixel point. By eliminating predefined anchor boxes, the hyperparameter design process of the anchor is removed, avoiding the situation where parameters affect network performance, while significantly reducing computational costs. Compared with the traditional one-stage algorithm or anchor-free algorithm, FCOS achieves better results in the task of object detection.

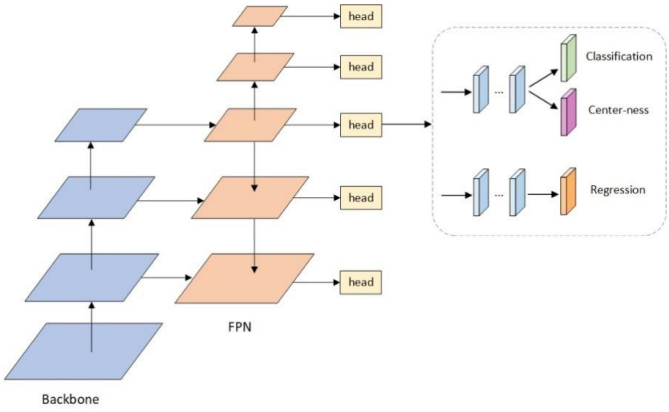

Figure 1. The network structure of FCOS.

The main innovation of FCOS is achieved in two parts. Firstly, it uses FPN for multi-scale prediction. The range of object sizes is restricted directly for each level and the distance between each location and the minimum bounding box of the object, i.e., l, t, r, b, is calculated.

These values are compared with the maximum regression distance of each level, to set positive and negative samples. The scale-corresponding multi-layer prediction can be realized based on the above operation, which largely enhances the prediction performance when bounding boxes overlap and avoids the problem of fuzzy matching.

Since the FCOS uses a pixel-by-pixel regression strategy, it produces a large number of low-quality predicted bounding boxes far from the centers of the object during training. Therefore, a branch parallel to the classification is added to predict the centeredness of the position, which describes the normalized distance from the position responsible for this object to its center. When testing, the centeredness values are multiplied with the corresponding classification scores to rank the bounding boxes and reduce the weight of the bounding boxes which are away from the center of the object. Eventually, non-maximal suppression (NMS) was used to remove bounding boxes that have low scores, significantly improving detection performance.

However, in complex scenes, FCOS will be affected by occlusion, which leads to its low detection accuracy. In addition, it cannot realize the task of instance segmentation, so this paper improves and optimizes it based on the FCOS model.

4. Methods

In this section, we improve the internal network of FCOS, aiming to achieve object detection as well as instance segmentation in a two-branch structure and to improve the experimental accuracy to some extent.

4.1. Sample-Fused FPN

To enhance the effectiveness of feature extraction, FPN connects high-level features to low-level features horizontally in a top-down order which both have different resolutions and semantic information. This operation allows feature maps at all scales to have rich semantic information, thus making independent predictions at different levels. Figure 2a illustrates the basic structure of the FPN, the FPN output p_i of the feature map ci corresponding to the backbone network can be expressed as Equations (1) and (2). Di stands for bottom-up backbone network, and U_i represents the bottom-up FPN operation (up-sampling and feature fusion).

$$c_i = D_i(c_{i1}) \tag{1}$$

$$p_i = U_i(c_i, p_{i+1}) \tag{2}$$

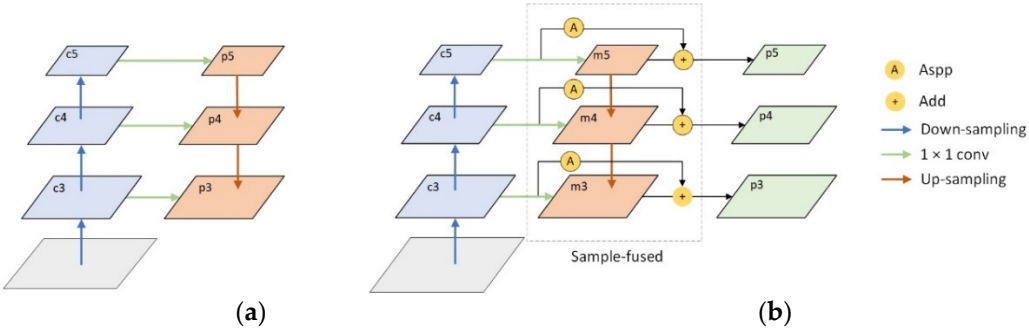

Figure 2. (**a**) The basic structure of the FPN; (**b**) sample-fused FPN.

However, in complex scenes, environmental factors can have an impact on the feature extraction of the object. So, in garbage detection, the variability in size between different garbage types requires the feature maps to contain semantic and spatial information at different scales. In this paper, we propose a new FPN structure, called sample-fused FPN, with the structure shown in Figure 2b. The ASPP module is introduced based on the original FPN structure. The computation flow of the feature map is as follows:

$$m_i = U_i(c_i, m_{i+1}) \tag{3}$$

$$a_i = A_i(c_i) \tag{4}$$

$$p_i = S_i(a_i, m_i) \tag{5}$$

m_i and p_I are similar which both represent the feature maps obtained after FPN. A_i is the ASPP module and the structure is shown in Figure 3. It uses three different operations for c_i, i.e., feature extraction based on dilated convolution, feature extraction based on 1×1 convolution, and global feature extraction based on a pooling operation of the image. It is equivalent to acquiring the contextual information of the image at multiple scales. Among them, DC1, DC2, and DC3 represent the dilated convolution with three different sampling rates, which are set to 6, 12, and 18, respectively. Then, we concatenate the five feature maps obtained by feature extraction and use 1×1 convolution to reduce the number of feature map channels so that the output is the ai. Si stands for feature fusion, which fuses the top-down output feature map with the feature map outputted by the ASPP module in an additional way to obtain pi in the end.

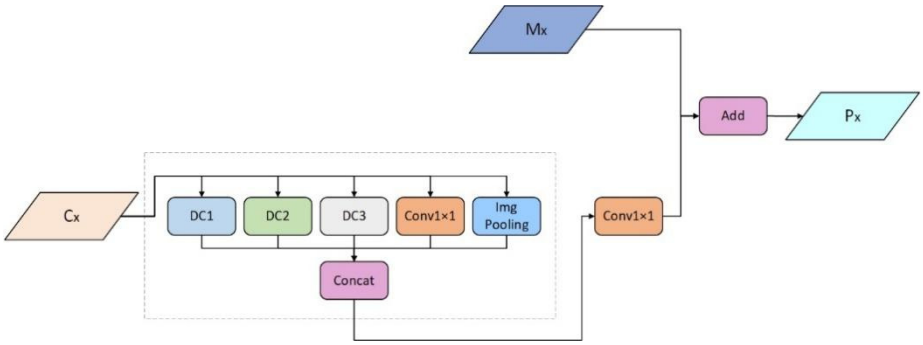

Figure 3. Structure of the ASPP module. This structure enables the acquisition of contextual information about the image at multiple scales.

4.2. Mask Head Based on Conditional Convolution

In conventional algorithms, for instance, segmentation, most of them use full convolutional networks to classify images on a pixel-by-pixel basis. Although the segmentation mask of the object is obtained in this way, the effect is relatively poor, especially in an environment with more occlusions around the object. Therefore, this paper improves the traditional mask branch by introducing the idea of conditional convolution and dilated convolution to increase the accuracy of instance segmentation. Figure 4 shows the improved mask branch.

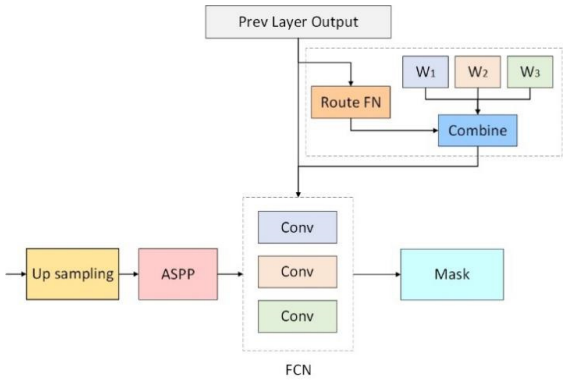

Figure 4. The mask head contains conditional convolution.

For each instance, a specific convolution kernel parameter is learned, and by replacing the standard convolution, it can increase the size and capacity of the model while maintaining efficient inference. The computation of parameters based on conditional convolution is shown in Equation (6).

$$\text{Output}(x) = \sigma((\alpha_1 \cdot W_1 + \ldots + \alpha_n \cdot W_n) * x) \quad (6)$$

As shown above, the parameters of the convolution kernel are transformed into a linear combination of n experts where $\alpha_1 \sim \alpha_n$ are the weighting coefficients learned by gradient descent. Subsequently, we assign these dynamically generated filters to the fully convolutional network. For an image with K instances, K different mask heads will be automatically generated, and the filter in each mask head contains the features of the corresponding instance. Therefore, in the instance segmentation task, it will only be triggered for the pixels of the corresponding instance, eventually producing the mask prediction of the instance. Compared with RoI, it can better represent irregular shapes. In addition, we add the ASPP module in front of the fully convolutional network. By combining a series of dilated convolutions, a sufficiently large receptive Field is obtained while fusing multi-scale information. As a result, the output features not only cover a large range of semantic information but also do information encoding in a very dense manner,

which can assist the fully convolutional network in pixel-by-pixel classification, to achieve high accuracy in the task of instance segmentation.

4.3. The Overall Structure of the Network

As can be seen from Figure 5, compared to the structure of FCOS, we have improved Neck by replacing the FPN with the SF-FPN constructed in this paper and introducing a Mask head based on conditional convolution. Based on the above modification, the network can be trained in an end-to-end form so that it can eventually achieve object detection and instance segmentation.

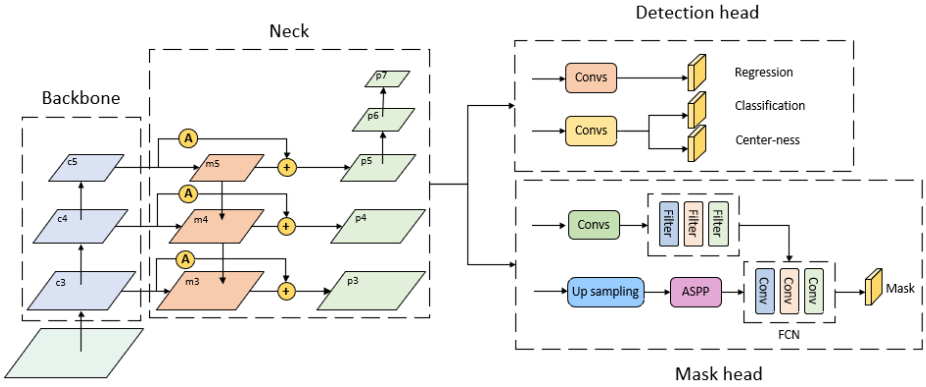

Figure 5. The improved network structure is based on FCOS.

First, we input an image that has the size of $H \times W \times 3$, and the features of the image are extracted based on ResNet. Then, the feature maps extracted from each layer of the backbone network are input into SF-FPN to optimize the feature information. Subsequently, we input the obtained feature maps into the Detection head and Mask head. The detection head aims at classifying and regressing bounding boxes and optimizing them by the centeredness function to achieve object detection. The Mask head is divided into two parts. In the first part, the corresponding network parameters are generated for each instance based on conditional convolution and combined into filters of the fully convolutional network. In the second part, up-sampling the input feature map and expanding it to one-eighth of the original image size is achieved. Finally, a fully convolutional network is used to classify the input feature maps pixel by pixel, generating the mask corresponding to each object.

4.4. Loss Function

The total loss function of the network model in this paper can be defined as:

$$L_{overall} = L_{detection} + \lambda L_{mask} \tag{7}$$

Since the detection branch is constructed based on FCOS, the focal loss [26] and the loss of Intersection over Union (IoU) are used for classification and regression, respectively. However, it is impossible to directly optimize the part which has no overlap and poorly discriminates different ways of alignment by using IoU as a loss. Therefore, this paper introduces Generalized Intersection over Union (GIoU) [27]:

$$GIoU = IoU - \frac{A^c - U}{A^c} \tag{8}$$

Among them, A^c represents the minimum enclosing box of the bounding box and ground truth. U is the combined area of the bounding box and ground truth. Because the GIoU introduces the concept of A^c, the regression can still be optimized when the bounding box and ground truth do not overlap. It can better reflect the overlap of the two boxes. In

general, GIoU retains the original features of IoU while weakening its drawbacks. The loss function of GIoU is defined as:

$$L_{GIoU} = 1 - GIoU \tag{9}$$

$L_{detection}$ and L_{mask} are defined as Equations (10) and (11):

$$L_{detection} = \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{x,y}, c_*^{x,y}) + \frac{\lambda}{N_{pos}} \sum_{x,y} o_{\{c_x^{x,y}>0\}} L_{reg}(t_{x,y}, t_*^{x,y}) \tag{10}$$

$$L_{mask}(\{\theta_{x,y}\}) = \frac{1}{N_{pos}} \sum_{x,y} o_{\{c_x^{x,y}>0\}} L_{dice}(FCN(F_{x,y}; \theta_{x,y}), M_*^{x,y}) \tag{11}$$

Since we use the original detection head of FCOS in this paper, we do not introduce its loss function. In the loss calculation of the mask head, o is the indicator function. Because of the use of the sigmoid function in the mask head to determine the front and back view of the pixel, it is a binary classification. The value is 1 when $c_x^{x,y} > 0$ and 0 otherwise. N_{pos} is the number of positive samples that are judged as foreground when $c_x^{x,y} > 0$. $F_{x,y}$, $\theta_{x,y}$ are the feature map and the parameters of the convolution at (x, y), respectively. $M_*^{x,y}$ is the mask of the corresponding instance generated at this position. Finally, the number of positive and negative samples is balanced by L_{dice}, similar to focal loss.

5. Experiments

5.1. Dataset

This paper uses the TACO dataset, which is an open image dataset of waste in the field. Figure 6 shows some of the images in this dataset. It contains 1500 photos of garbage taken in various environments which have 4784 examples, and the environment in which the garbage is located is divided into seven categories, and the garbage category is divided into sixty categories. Moreover, it provides annotations corresponding to the objects in the images that are defined in the form of a COCO dataset. These annotations consist of the information on the bounding box, the segmentation mask, and the category. Based on this dataset, the training and testing of algorithms in this application scenario can be implemented.

Figure 6. TACO dataset.

5.2. Evaluation Metrics

To evaluate the effectiveness of this algorithm in garbage detection and instance segmentation, we use the AP as an evaluation metric.

The AP (average precision) is the average value of precision at different recalls.

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

Equations (12) and (13) calculate the precision and recall, respectively. Among them, True Positive (TP) represents the samples that are predicted to be correct and positive.

False Positive (FP) represents the samples that are predicted to be positive but are actually negative. On the contrary, False Negative (FN) represents the samples that are predicted to be negative but are actually positive. Ultimately, the AP can be calculated by Equation (14).

$$AP = \int_0^1 P(R)dR \tag{14}$$

Since the TACO dataset contains multiple classes of objects, we use mAP to evaluate the overall effectiveness of the algorithm on the dataset. It is the average value of AP for different categories. The mAP is represented as Equation (15). N is the number of categories that exist in the dataset.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{15}$$

5.3. Training

The configuration of the experimental platform which is applied in this paper is shown in Table 1.

Table 1. Configuration parameters.

Devices	Configuration
Operating system	Ubuntu18.04
Processor	Intel(R) Xeon(R)Silver 4110 CPU @2.10 GHz
GPU	GeForce RTX 2080Ti
GPU accelerator	CUDA 10.2
Frame	Pytorch
Compilers	Vscode
Scripting language	Python3.8

Since the TACO dataset has only 1500 images at this stage, we improve the sample quality by data enhancement such as flipping, rotating, and random cropping of the original images. Finally, the dataset is divided into ten equal parts and the model is trained by using the ten-fold cross-validation method. The detailed operation is shown in Figure 7.

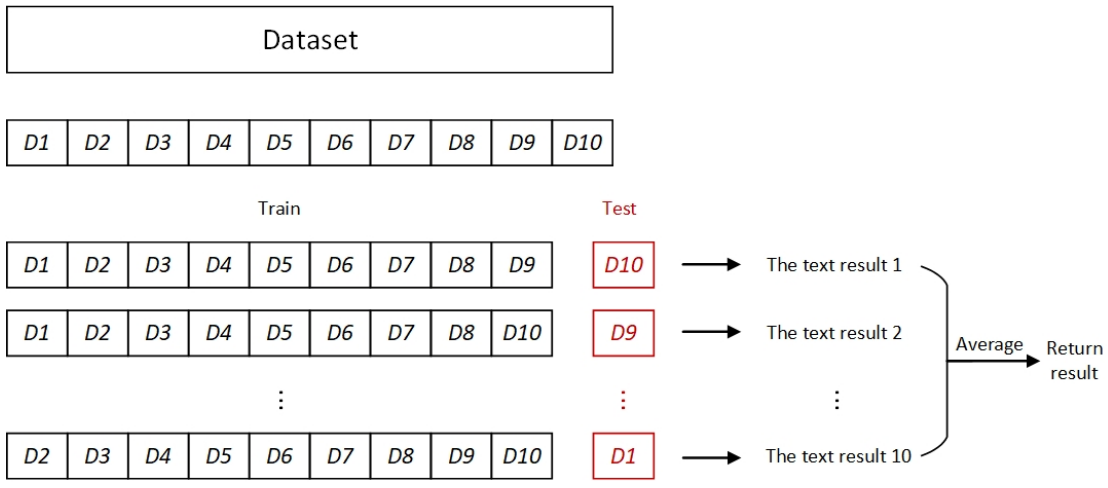

Figure 7. Training the network based on the ten-fold cross-validation method.

In the basic structure of the network, we used ResNet-101 [28] as the backbone network and initialized it with weights pre-trained on ImageNet to give the model some feature extraction capability that reduces the training time. During the training phase, the batch size is set to 12 and the initial learning rate is 0.01. On this basis, the network parameters are optimized by using stochastic gradient descent (SGD), where the momentum, as well

as decay weights, are 0.9 and 0.0001, respectively. Finally, the number of iterations is set to 100 k. When the values of iteration are 60 k and 80 k, the learning rate becomes one-tenth of the original automatically.

5.4. Visualization of Test Results

Figure 8 shows the effect of the algorithm proposed in this paper for object detection and instance segmentation on the TACO dataset. It can be seen that this algorithm can be better applied to complex scenes. It is not only able to achieve accurate object detection, the effect of generating a mask for different instances is relatively good.

Figure 8. The effect display of object detection and instance segmentation.

5.5. Performance Comparison of Different Algorithms

To scientifically demonstrate the effectiveness of the algorithm proposed in this paper applied to complex scenarios, we trained other algorithms simultaneously on the TACO dataset and compared their performance on the dataset of the test. The results are shown below.

As can be seen from Table 2, the algorithm proposed in this paper achieves better results on garbage detection in complex scenarios. Compared with the Faster R-CNN, the mAP index of our method is improved by 8.6. Compared with the Cascade R-CNN [29], the mAP index of our method is improved by 3.0. Compared with the FCOS, which is more frequently used at this stage, our method in this paper does not lose an advantage in accuracy. Its mAP index increased by 1.5. Since the method in this paper is improved based on FCOS, the mAP_s index is increased by 1.6, which is enough to prove that the improvement of the FPN structure can bring some effectiveness. It enhances the capability of feature extraction so that the feature maps at each scale get sufficient feature information and the features of small objects are more easily perceived, to locate and classify the objects effectively. Overall, our proposed algorithm in this paper can handle the task of object detection in complex scenes and achieve better detection accuracy.

Table 3 shows the final results obtained for each algorithm tested on the instance segmentation task. Since the FCOS algorithm can only implement the target detection task, we add the mask branch to its structure so that it can complete the instance segmentation simultaneously. It can be seen that the algorithm proposed in this paper also presents

better results on this task. Compared with the Mask R-CNN, the mAP index of our method is improved by 4.6. Compared with the MS R-CNN [30], the mAP index of our method is improved by 3.8. Compared with the recently proposed algorithm, SOLO [31], the mAP index of our method is improved by 2.2. In addition to this, for different sizes of the object, its mAP_s index, mAP_m index, and mAP_l index all improved by 1.8, 2.7, and 1.5, respectively. Therefore, the mask branch introduced in this paper not only helps the network to accomplish the basic instance segmentation task but also achieves better results compared with other algorithms.

Table 2. Performance comparison of algorithms on object detection.

Method	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Faster R-CNN	74.1	85.0	81.4	18.5	68.6	84.3
Cascade R-CNN	79.7	84.6	82.8	24.1	67.2	91.3
FCOS	81.2	86.1	83.5	23.9	69.4	92.0
Our method	82.7	86.8	85.1	25.7	71.2	92.3

Table 3. Performance comparison of algorithms on instance segmentation.

Method	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Mask R-CNN	67.8	83.3	74.4	4.5	61.2	75.8
MS R-CNN	68.6	82.4	75.4	4.4	61.1	76.4
SOLO	70.2	83.5	76.7	5.1	62.7	78.1
Our method	72.4	84.9	78.1	6.9	65.4	79.6

6. Conclusions and Outlook

As the level of materials required for human activities increases, it causes an increasingly serious situation of environmental pollution. Therefore, environmental protection has become a key research element to maintain stable social development at this stage.

In this paper, we propose a garbage detection algorithm that can be applied in complex environments according to an intelligent garbage cleaning task. First, to effectively reduce the impact of garbage size and environmental factors on the detection effect, an SF-FPN is constructed based on the dilated convolution, which enriches the semantic information of the feature map and improves the detection accuracy through feature extraction at multiple scales. Then, a mask branch based on conditional convolution is introduced to dynamically generate parameters for each object, resulting in the instance-level prediction of pixels. Finally, the network is trained end-to-end. The results show that the proposed algorithm is better than the current algorithm in garbage detection and instance segmentation tasks and achieves accurate detection and instance segmentation of garbage in complex scenarios. It provides an algorithmic basis for the subsequent automated cleaning of terrestrial waste by robots.

However, in the current work, there are still certain shortcomings that need to be improved. For example, the number of images in the TACO dataset is small. Although our algorithm has improved the accuracy compared to other algorithms, the overall accuracy of small object detection is low. In addition, how to optimize the network and reduce its computing resources is also a research content that needs attention. Finally, we will consider attaching the algorithm to a device to achieve practical application.

Author Contributions: Conceptualization, H.D. and D.E.; methodology, F.L. and B.M.; validation, Y.C. writing—review and editing, H.D. and D.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partially supported by grants from the National Natural Science Foundation of China #72174172, #U1811461 and #71774134; also supported by "the Fundamental Research Funds for the Central Universities", Southwest Minzu University (ZYN2022013).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Pedro F Proença, Pedro Simões. TACO: Trash Annotations in Context for Litter Detection. arXiv:2003.06975 [19]. GitHub—pedropro/TACO: Trash Annotations in Context Dataset Toolkit.

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

References

1. Li, G.; Kou, G.; Peng, Y. Heterogeneous Large-Scale Group Decision Making Using Fuzzy Cluster Analysis and Its Application to Emergency Response Plan Selection. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *52*, 3391–3403. [CrossRef]
2. Li, G.; Kou, G.; Li, Y.; Peng, Y. A group decision making approach for supplier selection with multi-period fuzzy information and opinion interaction among decision makers. *J. Oper. Res. Soc.* **2021**, *73*, 855–868. [CrossRef]
3. Li, Y.; Kou, G.; Li, G.; Peng, Y. Consensus reaching process in large-scale group decision making based on bounded confidence and social network. *Eur. J. Oper. Res.* **2022**, *303*, 790–802. [CrossRef]
4. Li, Y.; Kou, G.; Li, G.; Hefni, M.A. Fuzzy multi-attribute information fusion approach for finance investment selection with the expert reliability. *Appl. Soft Comput.* **2022**, *126*, 109270. [CrossRef]
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
6. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. [CrossRef]
7. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
8. Bochkovskiy, A.; Wang, C.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [CrossRef]
10. 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. [CrossRef]
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
12. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
15. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
16. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
17. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Advances in Neural Information Processing Systems, Vancouver, Canada, 8–14 December 2019; Volume 32.
18. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
19. Proença, P.F.; Simões, P. TACO: Trash Annotations in Context for Litter Detection. *arXiv* **2020**, arXiv:2003.06975.
20. Lee, S.-H.; Yeh, C.-H.; Hou, T.-W.; Yang, C.-S. A Lightweight Neural Network Based on AlexNet-SSD Model for Garbage Detection. In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference, Guangzhou, China, 22–24 June 2019; pp. 274–278. [CrossRef]
21. Ma, W.; Wang, X.; Yu, J. A Lightweight Feature Fusion Single Shot Multibox Detector for Garbage Detection. *IEEE Access* **2020**, *8*, 188577–188586. [CrossRef]
22. Cao, L.; Xiang, W. Application of Convolutional Neural Network Based on Transfer Learning for Garbage Classification. In Proceedings of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 12–14 June 2020; pp. 1032–1036. [CrossRef]
23. Kang, Z.; Yang, J.; Li, G.; Zhang, Z. An Automatic Garbage Classification System Based on Deep Learning. *IEEE Access* **2020**, *8*, 140019–140029. [CrossRef]
24. Shi, C.; Xia, R.; Wang, L. A Novel Multi-Branch Channel Expansion Network for Garbage Image Classification. *IEEE Access* **2020**, *8*, 154436–154452. [CrossRef]

25. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9627–9636. [CrossRef]
26. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
27. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over union: A metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
29. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
30. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6402–6411.
31. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*; Lecture Notes in Computer Science Book Series; Springer: Cham, Switzerland, 2020; pp. 649–665.