# Litter Detection from Digital Images Using Deep Learning

Jianfeng LIU[1], Chen PAN[1], Wei Qi YAN[2*],
(email: 1052431445@qq.com; pc916@cjlu.edu.cn)
[1] China Jiliang University, Hangzhou, China
[2] Auckland University of Technology, Auckland, New Zealand
(email: weiqi.yan@aut.ac.nz)

**Abstract.** In order to achieve automatically litter detection in residential area, machine vision has been applied to monitor environment of surveillance. Based on our observations and comparative analysis of the current algorithms, we propose an improved object detection method based on Faster R-CNN algorithm and achieve more than 98% accuracy of litter detection in surveillance. Through our observations, most of litters are small objects, we apply Feature Pyramid Network (FPN) to Faster R-CNN and optimize it by merging different layers by using multiply operate. Besides, we replace cross-entropy loss function with focal loss function to solve the problem of anchor imbalance by using Region Proposal Network (RPN) and offer attention module through RPN to feedback the whole network. We collected more than 8,000 labeled images from our surveillance videos for model training. Our experiments show that the improved Faster R-CNN achieves a satisfied performance in real scene.

**Keywords:** Litter detection, object detection, FPN, attention module

## 1 Introduction

Nowadays, we are use of computer vision to resolve the problem of garbage classification almost everywhere. High-definition cameras, high-speed 5G networks, and powerful computers enable this technology. However, there are still a number of challenges in the research of litter detection. Small size objects in picture, luminance changes, moving occlusions and public facilities in ground all cause unexpected problems for litter detection.

At present, a great deal of well-performed algorithms based on deep learning are proposed to solve object detection problem [32,33,34]. The representative one-stage network has SSD and YOLO [14, 20] series, the two-stage network has Faster R-CNN [22] and Mask R-CNN [5]. Because of its direct convolutional layer regression and classification, the one-stage network brings speed advantages. On the other hand, because the network has more steps such as proposal box extraction, its detection accuracy has improved tremendously. However, in litter detection, usually litter is a

* Corresponding author: Wei Qi Yan (email: weiqi.yan@aut.ac.nz)

small object before a camera, therefore a special method is needed to detect them. Although SSD and YOLO [14, 20] are relatively fast, there are limitations while dealing with small object, especially YOLO models. At the same time, Mask R-CNN [5] is an instance segmentation [23] network, and an edge detection [17] branch is added to the classification and regression tasks. It has the highest detection accuracy and loses the detection speed. Faster R-CNN [22] has been a classic object detection model till now, with satisfactory detection accuracy and appropriate detection speed. The proposed methods automatically extract visual features and have better generalization capabilities.

Overall, our contributions of this paper are listed as follows:

- We apply FPN [12] to Faster R-CNN [22] and conduct an improved modification with the FPN structure using multiply operate to merge each layer.
- We replace the CE loss [3] to focal loss [13] in Region Proposal Network (RPN) [22]. Besides, we implement multiple anchor scales and select the best group in our litter scenes.
- By using RPN [22] as anchor and attention module, the attention mechanism [30] can effectively react to the overall network. This makes full use of RPN's foreground and background discrimination capabilities at a small cost.
- We collect a litter dataset with 8,000 labeled images from surveillance videos of real scenes and achieve more than 98% accuracy in litter detection.

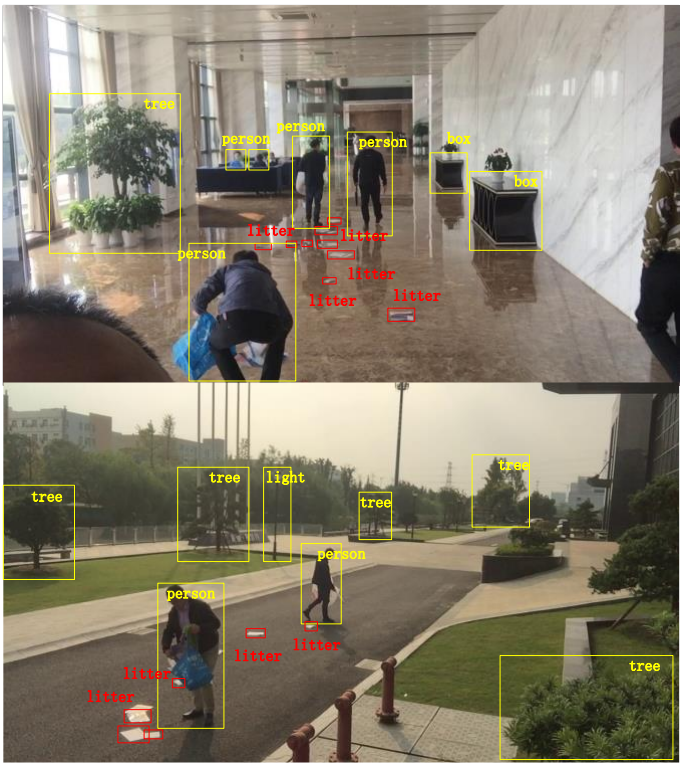

Fig.1 Litters in real scenes, red rectangles represent the detected litters, yellow boxes show the normal visual objects.

## 2 Related Work

Litter detection has achieved great results. Mittal et al. [18] proposed a neural network based on AlexNet [9] to detect the region of litter dump, achieved 87.69% accuracy and employed it to mobile phone applications. AlexNet [9] has conducted experiments to get the better performance compared to traditional image processing but it is still an early convolutional neural network and did not get the best results. Lee et al. [10] proposed a modified Single Shot MultiBox Detector (SSD) [14] with lightweight model. The backbone network of SSD was created based on AlexNet [9], but it is to deal with large-scale recyclable litter at close range, which is not suitable for small-scale litter scenes under long-range cameras. Wang et al. [29] has implemented ResNet [6] to replace Faster R-CNN [22] of VGG [24] network to detect garbage. He also proposed a strategy to merge litter data with multiple scenarios for training, and achieved 89% accuracy in the final results.

Visual object detection has two major tasks: Positioning and classification. The positioning is to represent the object position as a boundary or bounding box; meanwhile, the classification is to predict the class of the given objects. Fast R-CNN [4] was applied to solve the problem of object detection. It was evolved from R-CNN series and an RPN network was proposed for extracting region proposal. RPN [22] is a lightweight and simple network, which conducts regression and classifications so as to obtain rough object candidate boxes [8]. Because its process can be accelerated by using GPU, it is much faster than selective search [28] method. After the RPN process, it takes use of pooling [22] to generate feature maps of the same size and send it to classification and regression tasks, respectively. In addition, Mask R-CNN [5] was from a fully convolutional neural network [16] based on Faster R-CNN [22] to add additional segmentation tasks [15], improve the Pooling to RoIAlign [4], make the positioning much accurate through bilinear interpolation. In the object detection task, the identified object is mostly with a medium size, while the litter detection in this task basically is related to the detection of small-size objects, and the size of the target is also affected by the distance from the cameras.

As shown in Fig. 1, before two digital cameras, through the label picture of man-made manual annotation, yellow boxes include a variety of detected objects, red boxes show the detected litters, we see the area of litters are smaller than that of general objects.

## 3 Our Methods

Most object detection tasks cannot work well for the detection of small-size visual objects. Therefore, in this paper, we compare various visual object detection frameworks and select Faster R-CNN algorithm for this project.

- The small-size objects may lose detailed information after a large number of convolution operations. Therefore, in this paper, feature pyramid structure combined with Faster R-CNN framework is introduced to solve the problem of feature loss by using scaling and merging.
- The FPN solves the problem of fusion, but in the process of fusion, modifying convolutional and sampling operations could improve the feature map and make small object have good output, which was verified in our experiments.
- The FPN structure has the problem of lacking top-level data sources, the SAM module can be employed to strengthen the top-level information sources.
- In Faster R-CNN, attention mechanism is applied to improve the final result without increasing the computing costs.

### 3.1 The Structure of Neural Network

As shown in Fig.2, we create our network based on Feature Pyramid Network (FPN) [12] and Faster R-CNN [22] which gets helpful U-shape frame designed by using top-down and bottom-up methods. Owing to the multiscale [2] feature of FPN, this network represents global information very well, especially for small objects in a given image. The whole process is to input an image through the FPN module and apply R-CNN for RoI pooling [4], regression, and classification.

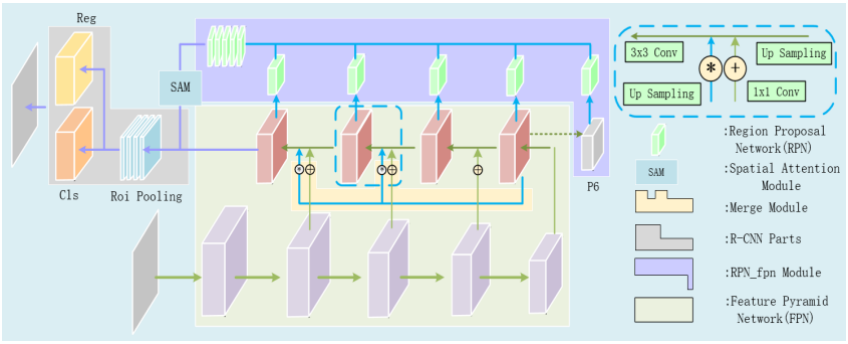


Fig.2 The pipeline of our proposed approach. At the top-right corner, the details of multiply merges are shown in this model.

### 3.2 FPN of Merge Module

As we all know, the focus of low-level features is on local areas with detailed information, the high-level features have larger receptive fields and better positioning information. Feature pyramid network is use of the information of each feature layer to make the final output result combined with multiple layers of the information. FPN take use of layer-by-layer to fuse high-level information [11] with low-level information, as shown in eq. (1),

$$f_i = \theta(f_j, f_i) = \tilde{n}(\eta_3(f_j)) + \eta_1(f_i)\eta_i \quad (1)$$

where $\eta_i$ represents i×i convolution layer and O refers to upsampling operation, fj stands for i-th feature map. We believe that in visual object detection, the operations will also gradually dilute high-level features and reduce the role of high-level features. Therefore, we propose a method for fusing the features of the highest layer with the features of each layer to achieve complementary information [25]. As shown in Fig. 2 (yellow region), the results show that this method effectively improves the detection results of Faster R-CNN [22]. In the fusion part, the features of this layer are firstly added and fused with the up-sampled features of the previous layer, then the results are continuously multiplied and fused with the features of the highest layer, as shown in the dotted box in the top right corner of Fig. 2.

### 3.3 Region Proposal Network of FPN

In this section, we introduce the RPN [22] of FPN [12]. Since the network structure becomes a bottom-up and top-down encoding and decoding structure, the RPN must have been changed relatively.

Firstly, after C5 is convoluted to get P5, P6 layer is obtained by downsampling with P5, P2, P3, P4, P5, and P6 are sent to the RPN layer for obtaining the regional proposals, respectively. The regional proposal obtained from the five feature maps is combined and selected to remove overlap, negative samples and small boxes. Finally, we output the class and bounding box of the proposal, as shown in Fig.3.

Region-of-Interest (RoI) pooling is used to extract features. RoI pooling of FPN is different from RoI pooling of Faster R-CNN. Formally, we set a RoI with width $w$ and height $h$ (the network of the input image) to the level $p_k$ of our feature pyramid.

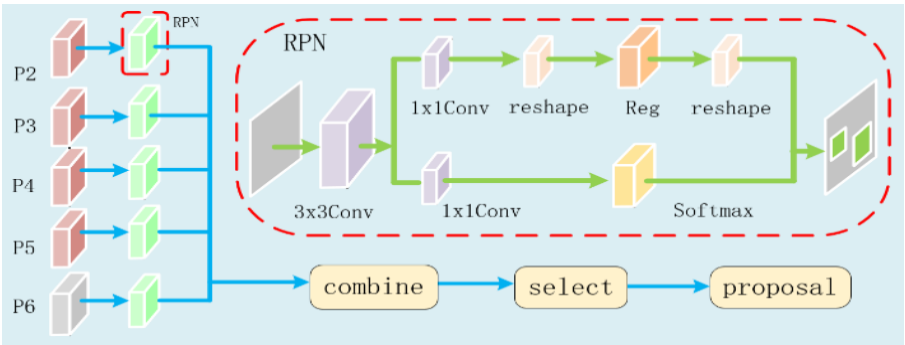


Fig.3 The RPN of FPN

### 3.4 Implement Focal Loss Function in Faster R-CNN

RPN [22] generates a large number of candidate boxes for Faster R-CNN, and these candidate boxes have the problems of unbalanced positive and negative samples and small boxes with low confidence. Focal loss [13] is very effective to solve this problem in one-stage object detection [26]. Cross-entropy loss function is often used as a loss function for classification problems, as shown in eq.(2),

$$CE(y') = -y\log y' - (1-y)\log(1-y') \tag{2}$$

The alpha coefficient can effectively avoid the uneven problem of the anchor box, the gamma index is used to solve the imbalance problem of the simple and difficult anchor box [31] as shown in equation (3).

$$FL(p_t) = \begin{cases} -\alpha_t(1-p_t)^{\gamma}\log p_t, & p_t = 1 \\ -(1-\alpha_t)p_t^{\gamma}\log(1-p_t), & p_t = 0 \end{cases} \tag{3}$$

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{4}$$

where for binary classification, $p_t \in [0,1]$ is the probability for the ground true class, $\alpha_t \in [0,1]$ is the re-weighting factor, $\gamma \geq 0$ is a hyper-parameter. In eq. (5), $\alpha_t$ is a scale factor, which is fallen in [0.6, 1], $\gamma$ is a power index.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{5}$$

The loss of RPN is composed of two parts. As shown in Eq. (5), the left half is the classification loss and the right half is the regression loss. $P_i$ and $t_i$ are the output results. $p_i^*, t_i^*$ is the corresponding label. In the subscript, *CLS* stands for classification and *Reg* stands for regression In this paper, focal loss was used to improve the classification loss of the left half to solve the problem of unequal positive and negative samples in the process of candidate box extraction.

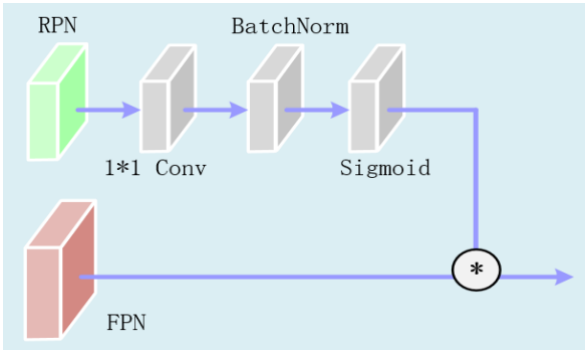


Fig.4 Spatial attention module

### 3.5 Spatial Attention Module

Throughout the process [19], we see that RPN made a distinction between foreground and background, though it was not very accurate. We make use of RPN by treating the results of RPN as an attention model to update the output of FPN. In this way, without additional computational burden, RPN is used to distinguish foreground and background regions [7] more effectively.

Firstly, the output characteristic graph of RPN is saved, then the characteristic graph of RPN is convolved by $1 \times 1$ to make the channel number as same as the channel number of FPN. The feature graph is processed by using batch normalization to make the data more regular. We go through the sigmoid operation to get the probability between 0 and 1. Finally, the characteristic graph of FPN is multiplied by the characteristic graph of RPN as shown in Fig.4.

## 4 Results Analysis

In our experiments, there were eight surveillance videos shown the process of littering, each video duration is from 3 to 5 minutes, the video resolution is 1280 ×720. Among them, traditional data enhancement methods such as rotation, mirroring, brightness adjustment, and adding noises have been applied to image enhancement so as to generate 8,000 images. We used 1,600 images as our test set. The remaining 6,400 images are split into a training set and a validation set.

In this paper, we took use of Faster R-CNN based on feature pyramid network, the backbone of Faster R-CNN is ResNet101. The scales of FPN anchor are 32, 64, 128, 256, and 512; the FPN feature strides are 4, 8, 16, 32, 64; anchor ratios are 0.5, 1, 2. Under the batch size 2, 0.001 learning rate, and 12 epochs, the pooling mode is RoIAlign, the decay step is set as five, decay rate is assigned as 0.1. For our hardware, we utilize a GPU of NVIDIA TITAN X and Intel I7 CPU.

In order to measure the performance of the target detection algorithm, the intersection ratio (IoU) represents the overlap rate of the bounding box, where *area* represents the area of the bounding box, *pred* is the predicted box, and *gt* is the ground truth, as shown in eq.(9),

$$IoU = \frac{Area(pred) \cap Area(gt)}{Area(pred) \cup Area(gt)} \tag{6}$$

If IoU is greater or equal to *x*, the class is *c*, where the total is all objects in the images as shown in eq. (10),

$$Precision_{c,iou=x} = \frac{N(IoU \geq x)_c}{N(Total)_c} \tag{7}$$

The average accuracy (AP) refers to the average value of the accuracy rate of multiple objects as shown in eq. (11),

$$AP_c = \frac{\sum Precision_c}{N(Total)_c} \tag{8}$$

In the top half of Table 1, Faster R-CNN surpasses YOLO v3 [20] and YOLO v4 [1]. The reason is that the object detected by this task is a small object, the RPN operation of two-stage network is more sensitive to small targets, while YOLOv3 and YOLOv4 are not sensitive to small objects. The FPN can greatly increase the accuracy of garbage detection, this indicates that multi-scale can well improve the accuracy of small objects. Other methods in this paper (multiplication fusion, focal loss and spatial attention model) all improve the detection accuracy to varying degrees. It can be found from the fifth and sixth lines that the effect of multiplication fusion is higher than that of addition fusion, the result has been significantly improved. The increase of focal loss and spatial attention respectively improved the detection accuracy, and the model integrated with all the methods achieved the best detection effect, demonstrating the effectiveness of the module.

Our dataset consists of 8 scenarios. Table 2 shows the detection results of multiple methods in various scenarios. Among them, the detection error rate of Scene 4, 7 and 8 is obviously higher than that of other scenes. The reason is that the ground in this scene is reflective from ceramic tiles, which affects the visual object detection. Besides, the camera is far away from the garbage, which leads to the small size of the garbage. Compared with different methods, the original Faster R-CNN has the worst effect, and the detection method using attention mechanism and multiplication fusion method is better than others.

Table 1: The result comparisons with multiple methods: FPN, MM, AM, FL and SA

| FPN | MM | AM | FL | SA | AP | FP/TP | Miss rate |
|---|---|---|---|---|---|---|---|
| YOLO v3 | | | | | 0.8021 | 1482/6451 | 0.43 |
| YOLO v4 | | | | | 0.8846 | 663/6928 | 0.26 |
| Faster RCNN | | | | | 0.921 | 300/7185 | 0.11 |
| √ | | | | | 0.9795 | 184/7540 | 0.04 |
| √ | √ | | | | 0.9811 | 162/7571 | 0.03 |
| √ | | √ | | | 0.9807 | 181/7562 | 0.04 |
| √ | √ | | √ | | 0.9818 | 159/7567 | 0.03 |
| √ | √ | | √ | √ | 0.9841 | 168/7587 | 0.03 |

Table 2: The FP/TP resultant comparisons with multiple methods in various scenes

| scenes | multimerge | attention | addmerge | FPN | faster |
|---|---|---|---|---|---|
| 1 | 7/311 | 6/312 | 10/310 | 6/311 | 5/310 |

| | | | | | |
|---|---|---|---|---|---|
| 2 | 12/816 | 7/816 | 5/818 | 7/817 | 11/816 |
| 3 | 3/669 | 4/669 | 5/668 | 5/668 | 40/639 |
| 4 | 32/1068 | 21/1072 | 36/1072 | 34/1070 | 57/1058 |
| 5 | 13/658 | 7/657 | 15/651 | 12/660 | 110/619 |
| 6 | 26/957 | 30/953 | 31/952 | 26/954 | 118/898 |
| 7 | 23/1249 | 23/1250 | 23/1248 | 24/1241 | 77/1146 |
| 8 | 61/1844 | 68/1848 | 56/1843 | 70/1839 | 305/1542 |

Figure 5 and Fig. 6 show the experimental results. We see that the proposed method was superior to the original Faster R-CNN and unmodified FPN network. In Fig. 5, the vertical axis shows the error rate. The error rate using attention mechanism and multiplication fusion is significantly lower than other methods, and it has been verified in multiple scenarios. Figure 6 shows the MAP, the attention mechanism achieved the highest score, which demonstrates the effectiveness of the method in this paper.

As shown in Fig. 7, Faster R-CNN and YOLOv4 achieved the lowest MAP value, while the methods in this paper are all improved, among which PR curve of the attention mechanism was the best.

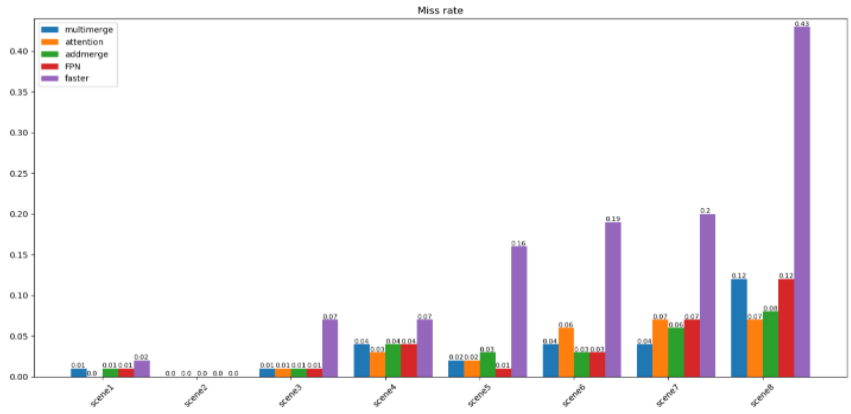


Fig.5 Miss rates for multiple scenes

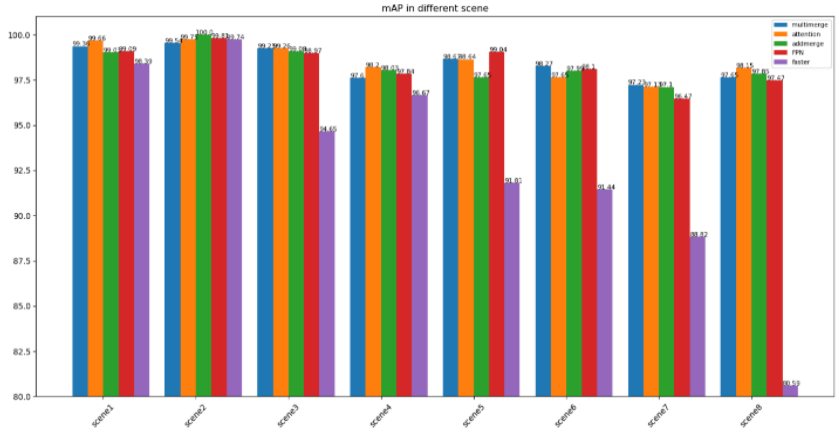


Fig.6 mAPs for multiple scenes

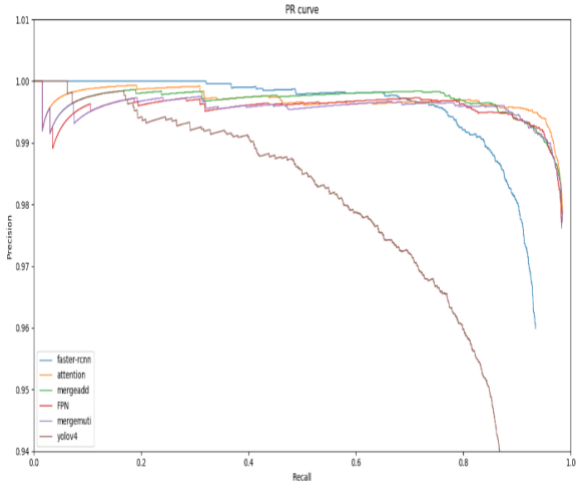


Fig. 7 The PR curve with multiple methods

## 5 Conclusion

In this paper, we implement a deep learning method for litter detection in digital surveillance. Faster R-CNN of FPN with attension model was employed as the core part, we collected a dataset for real scene of littering with 8,000 video frames. A new structure based on deep learning models was proposed. We apply focal loss to the RPN so as to solve the problem of the imbalance of the anchor box in the scenario where the actual garbage is a small object, using the FPN network of the multiplication fusion mechanism to handle small object detection. Spatial attention module is given to feedback foreground and background feature. Our experimental results show that in a real environment, this method can handle garbage detection tasks in various scenarios well, and achieve an accurate recognition rate of over 98%.

**Conflict of Interest:** The authors declare that they have no conflict of interest.

## References


[1] Bochkovskiy, A., Wang, C. Y., & Liao, H M.: YOLOv4: Optimal speed and accuracy of object detection. IEEE CVPR (2020)

[2] Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. European Conf erence on Computer Vision. 354-370 (2016)

[3] De Boer, P. T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y.: A tutorial on the cross-entropy method. Annals of Operations Research, 134 (1), 19-67 (2005)

[4] Girshick, R.: Fast R-CNN. International Conference on Computer Vision (ICCV). Santiago, 2015, pp. 1440-1448 (2015)

[5] He, K., Gkioxari, G., Dollár, P., & Girshick, R.: Mask R-CNN. IEEE International Conference on Computer Vision, pp. 2961-2969 (2017)

[6] He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognitio n, pp. 770-778 (2016)

[7] Heikkila, M., & Pietikainen, M.: A texture-based method for modeling the background and detecting moving objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4), 657-662 (2006)

[8] Kong, T., Yao, A., Chen, Y., & Sun, F.: HyperNet: Towards accurate region proposal generation and joint object detection. IEEE Conference on Computer Vision and Pattern Recognition (2016)

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Neural Information Processing Sy stems (2012)

[10] Lee, S. H., Yeh, C. H., Hou, T. W., & Yang, C. S. (2019). A lightweigh t neural network based on AlexNet-SSD model for garbage detection. High Performance Computing and Cluster Technologies Conference (pp. 27 4-278).

[11] Li, L., Su, H., Feifei, L., & Xing, E. P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. Neural Information Processing Systems (2010)

[12] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S: Feature pyramid networks for object detection. IEEE CVPR, pp. 2117-21 25 (2017)

[13] Lin, T., Goyal, P., Girshick, R., He, K., & Dollar, P.: Focal loss for dense object detection. IEEE ICCV (2017)

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & B erg, A. C.: SSD: Single shot multibox detector. European Conference on Computer Vision, pp. 21-37 (2016)

[15] Long, J., Shelhamer, E., & Darrell, T.: Fully convolutional networks for semantic segmentation. International Conference on Computer Vision and P

attern Recognition (2015)
[16] Ma, X., Chen, Z., & Zhang, J.: Fully convolutional network with cluster for semantic segmentation. International Conference on AMME (2018)
[17] Marr, D., & Hildreth, E. C.: Theory of edge detection. Proceedings of The Royal Society B: Biological Sciences, 207(1167), 187-217 (1980)
[18] Mittal, G., Yagnik, K. B., Garg, M., & Krishnan, N. C.: SpotGarbage: Smartphone App to detect garbage using deep learning. ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 940-945 (2016)
[19] Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., & Sun, J.: ThunderNet: Towards real-time generic object detection. International Conference on Computer Vision and Pattern Recognition (2019)
[20] Redmon, J., & Farhadi, A.: YOLOv3: An incremental improvement. arXiv: 1804.02767 (2018)
[21] Redmon, J., & Farhadi, A.: YOLO9000: better, faster, stronger. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263-7271 (2017)
[22] Ren, S., He, K., Girshick, R., & Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, pp. 91-99 (2015)
[23] Romeraparedes, B., & Torr, P. H.: Recurrent instance segmentation. European Conference on Computer Vision, 312-329 (2016)
[24] Simonyan, K., & Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
[25] Thakare, B. S., & Dube, M. R.: New approach for model merging and transformation. International Conference on Computer Communication and Informatics (2012)
[26] Tian, Z., Shen, C., Chen, H., & He, T.: FCOS: Fully convolutional one-stage object detection. IEEE Conference on Computer Vision and Pattern Recognition (2019)
[27] Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., & Liang, J.: Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. IEEE Conference on Computer Vision and Patt ern Recognition, pp. 7340-7351 (2017)
[28] Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W.: Selective search for object recognition. International Journal of Computer V ision, 104(2), 154-171 (2013)
[29] Wang, Y., & Zhang, X.: Autonomous garbage detection for intelligent urban management. MATEC Web of Conferences, Vol. 232, pp. 01056. E DP Sciences (2018)
[30] Xie, Y., & Chen, Y.: Object tracking based on spatial attention mechanism. Chinese Control Conference (2019)
[31] Zhong, Y., Wang. J., Peng, J., and Zhang, L.: Anchor box optimization for object detection, IEEE Winter Conference on Applications of Computer Vision, pp. 1275-1283 (2020)
[32] Yan, W.: Introduction to Intelligent Surveillance. Springer (2019)
[33] Shen, D., Xin, C., Nguyen, M., Yan, W.: Flame detection using deep

learning IEEE ICCAR (2018)
[34] Yan, W.: Computational Methods for Deep Learning. Springer (2021)

**Appendix**: The notation or abbreviation table for the symbols

| | |
|---|---|
| CE Loss | Cross-Entropy Loss |
| FPN | Feature Pyramid Network |
| R-CNN | Region-Based Convolutional Neural Network |
| RoI | Region of Interest |
| RPN | Region Proposal Network |
| SSD | Single Shot MultiBox Detector |
| VGG | Very Deep Convolutional Networks |
| YOLO | You Only Look Once |