Op-PSA: an Instance Segmentation Model for Occlusion of Garbage

Sheng Yu and Fei Ye

August 15, 2023

Op-PSA: an instance segmentation model for occlusion of garbage

Sheng Yu[1] and Fei Ye[2]

[1] Beijing University Of Technology, 100 Pingleyuan, Chaoyang District, Beijing, China
[2] Jilin Jianzhu University, 5088 Xincheng Street, Jingyue Development Zone, Changchun City, Jilin Province, China
lncs@springer.com

Abstract. With the increasing emphasis on green development, garbage classification has become one of the important elements of green development. However, in scenarios where garbage stacking occurs, the task of segmenting highly overlapping objects is difficult because the bottom garbage is in an obscured state and its contours and obscured boundaries are usually difficult to distinguish. In this paper, we propose an Op-PSA model, which uses the HTC model as the baseline model and improves the modeling method of backbone network and model interest region using attention model and occlusion perception model. The Op-PSA model constructs the image as two overlapping layers and uses the two-layer structure to explicitly model the occluded and occluded objects, so that the boundaries of the occluded and occluded objects are naturally decoupled, and their interactions are considered in the mask regression. It is experimentally verified that the model can effectively detect the masked garbage and improve the detection accuracy of the masked garbage.

Keywords: Instance segmentation, Garbage detection, Attention model, Occlusion recognition.

1 Introduction

Nowadays, garbage detection on urban streets often relies on manual sorting and recycling, which makes this task time-consuming and laborious [1]. Because garbage is often stacked together in the real world, which leads to the phenomenon of garbage being blocked, and the accuracy of detection and recognition of blocked garbage is seriously affected by the problem of incomplete information and fuzzy boundary information [2-3]. How to accurately detect and recognize blocked garbage is an important problem of Waste sorting, which has important research significance [4].

Therefore, this study investigates high-precision image segmentation for garbage with difficult feature extraction in the presence of occlusion, and provides a new solution for the optimization of domestic garbage detection methods.

In this paper, we use an instance segmentation technique based on the attention model and the occlusion perception model to solve the above problem. First, the attention model enables the neural network model to give different attention to different parts of the input data by simulating the attention allocation of the human brain, which in turn improves the detection quality of spam instance segmentation. Secondly, occlusion perception refers to modeling the region of interest in an image as two overlapping layers, with the upper layer detecting the occluder object and the lower layer inferring the partially occludee target object, thus deconstructing the boundary between the occluder object and the occludee object, and facilitating subsequent instance segmentation detection. Finally, we implement an instance segmentation model that can effectively improve the accuracy of spam detection.

2 Related Work

2.1 Garbage instance segmentation detection

Recently, the study of intelligent classification, detection, and segmentation of garbage using computer vision techniques has attracted a great deal of interest from researchers. The research of the garbage detection method based on instance segmentation is presented below. Xu [5] designed an algorithmic model applicable to street garbage recognition and detection, which is improved from a general sample segmentation model YOLACT. To address the problem of high hardware deployment requirements for garbage convolutional neural networks, Wang [6] proposed a simple and feasible garbage object segmentation algorithm based on RGBD features and SVM.

The aforementioned studies have been conducted to improve the detection speed of garbage, enhance the detection of garbage in the water context, and increase the utilization of spatial prior information on garbage. Existing research methods have improved the garbage detection model from different perspectives, but have not addressed the difficulty of detection in the case of occlusion caused by garbage accumulation.

2.2 Garbage instance segmentation detection

The detection of occluded objects has always been a hot and difficult research area in computer vision. In the detection of garbage, the garbage to be detected is often occluded or mutually occluded, and solving the problem of garbage occlusion can effectively improve the detection of garbage ground.

Rajaei [7] demonstrated the important role of the repetitive process of object recognition under occlusion conditions. Tian [8] et al. proposed the DeepParts model to improve the detector's detection for occluded pedestrian performance. Chu [9] et al. changed the proposed frame of the network based on the FPN network through modules such as EMD Loss to be able to predict multiple targets. To address the limitation problem of the current loss function for the occluded population, Wang [10] et al. proposed a bounding box regression loss function Repulsion Loss for the occluded population.

It is difficult to segment instances of unnatural garbage shapes or occluded garbage due to occlusion when garbage is stacked in occurrence and there is no significant difference between object contours and occlusion boundaries.

3 Op-PSA Model

To address the problem that obscured garbage is not easy to detect, this paper carries out research on the segmentation of obscured garbage instances based on the attention model.

3.1 Network Overview

Firstly, to solve the feature extraction problem, the relationship between channels is studied on the above basis, and the pyramidal squeezed attention to structure is introduced in the feature extraction network of the hybrid task cascade model; secendly, according to the occlusion relationship of the occlusion garbage, two overlapping layers are introduced to increase the occlusion relationship between objects for the construction process of the region of interest of the hybrid task cascade model. The general architecture of the model is shown in Figure 1.

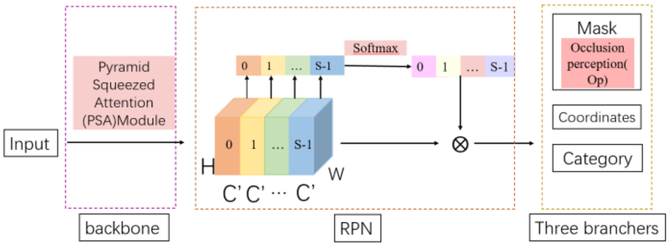

Fig. 1. Model Architecture diagram.

3.2 Pyramid Squeezed Attention (PSA)Module

Since the shape features of garbage are not easily extracted, this paper improves the feature extraction network of the hybrid task cascade model by considering the intrinsic connection between channels through the pyramid-squeezed attention structure Pyramid Squeezed Attention and automatically generates the weight of each feature channel according to the importance of the features using deep learning methods. On top of this, features with higher weights are enhanced, and conversely suppressed for features with lower weights.

The motivation for conducting this work is to build more efficient and effective channel attention mechanisms. To this end, a new Pyramid Squeezed Attention Module is proposed. As shown in Figure 2, the module is implemented in four main steps. First, the channel-level multiscale feature maps are obtained by implementing the proposed Pyramid Squeezed Attention Module. Second, the attention of the feature maps at different scales is extracted using the channel attention module to obtain the attention vector in the channel direction. Third, the channel attention vectors are recalibrated using Softmax to obtain the recalibration weights of the multi-scale channels. Fourth, the element-by-element product operation is applied to the recalibrated weights and the corresponding feature maps. Finally, as an output, a fine feature map containing richer multi-scale feature information is obtained.

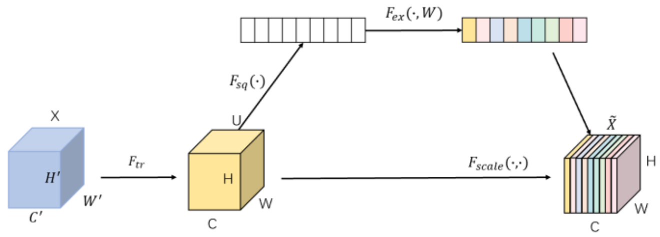

Fig. 2. Pyramid Squeezed Attention (PSA)Module.

As shown in figure 3, the basic operator that implements multiscale feature extraction in the proposed pyramidal squeezed attention model is the adaptive depth module, and the model extracts the spatial information of the input feature map in a multi-branch manner, with each branch having an input channel dimension of C.

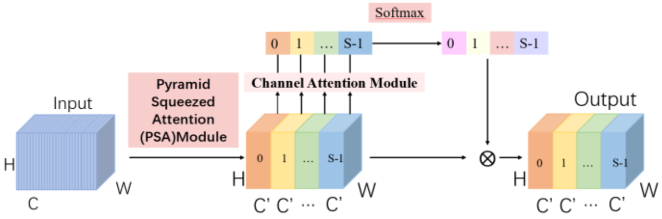

Fig. 3. Structure diagram of the pyramid squeeze attention module.

By doing so, the model can obtain richer information about the location of the input tensor and process it in a parallel manner at multiple scales. As a result, feature maps containing a single type of kernel can be obtained. Accordingly, different spatial resolutions and depths can be generated by using multi-scale convolutional kernels in a pyramidal structure. By compressing the channel dimensions of the input tensor, the spatial information at different scales on each channel feature map can be efficiently extracted. Finally, each feature map of different scales F_i has a common channel dimension $C' = \frac{C}{S}$ and i = 0, 1, ···, S - 1. At this point C should be divisible by S. For each branch, it learns multi-scale spatial information independently and builds cross-channel interactions in a local way. However, as the kernel size increases, the number of parameters increases significantly. In order to handle the input tensor at different kernel scales without increasing the computational effort, a group convolution method is introduced and applied to the convolution kernel. In addition, a new criterion is designed in this paper to select the group size without increasing the number of parameters. The relationship between multi-scale kernel size and group size can be written as

$$G = 2^{\frac{K-1}{2}}, \tag{1}$$

where the quantity k is the nuclear size and G is the group size. The above equations have been confirmed by ablation experiments, especially when k × k = 3 × 3 and G = 1. Finally, the multi-scale feature map generation function is given by the following eqution:

$$F_i = Conv(k_i \times k_i, G_i)(X) \quad i = 0,1,2,\cdots,S-1, \tag{2}$$

where the ith kernel size k_i = 2×（i +1）+1, the ith group size G_i = $2^{\frac{k_i-1}{2}}$, and F_i ∈ $R^{C' \times H\times W}$ denote the feature maps at different scales. The whole multi-scale preprocessed feature map can be obtained by cascading as

$$F = Cat([F_0, F_1,\cdots, F_{S-1}]). \tag{3}$$

F ∈ $R^{C\times H\times W}$is the obtained multiscale feature map. The attention weight vectors at different scales are obtained by extracting the channel attention weight information from the multi-scale preprocessed feature maps. The adaptive depth module is used to obtain the attention weights from the input feature maps at different scales. By doing so, the pyramid-squeezed attention module in this paper can fuse contextual information at different scales and generate better pixel-level attention for high-level feature maps. To achieve the interaction of attention information, the cross-dimensional vectors are fused without destroying the original channel attention vectors. And thus the entire multi-scale channel attention vector is obtained in a cascaded manner as

$$Z = Z_0 \oplus Z_1 \oplus \cdots \oplus Z_{S-1}, \tag{4}$$

where ⊕ is the concat operator, Z_i is the attention value from F_i, and Z is the multiscale attention weight vector. Soft attention is used across channels to adaptively select different spatial scales, which is guided by the compact feature descriptor Z_i. The soft assignment weights are given by the following equation:

$$att_i = Softmax(Z_i) = \frac{\exp(Z_i)}{\sum_{i=0}^{S-1}\exp(Z_i)}, \tag{5}$$

where Softmax is used to obtain a rescaled weight att_i of the multiscale channel that contains all the location information on the space and the attention weights in the channel. By doing so, the interaction of local and global channel attention is achieved. Next, the feature recalibrated channel attentions are fused and stitched together to obtain the whole channel attention vector as

$$att = att_0 \oplus att_1 \oplus \cdots \oplus att_{S-1}, \tag{6}$$

where att denotes the multi-scale channel weights after the attention interaction. Then, in this paper, the recalibrated weights of the multi-scale channel attention att_i are multiplied with the feature maps of the corresponding scales F_i as

$$Y_i = F_i \odot att_i \quad i = 1,2,3,\cdots,S-1, \tag{7}$$

where ⊙ denotes the channel multiplication and Y_i denotes the feature map after obtaining the multi-scale channel attention weights. The splicing operator is more efficient than the summation operator because it can keep the feature representation intact without destroying the information of the original feature map. In summary, the process of obtaining the refinement output can be written as

$$Out = Cat([Y_0, Y_1, \cdots, Y_{S-1}]). \tag{8}$$

As shown in the above analysis, the pyramid-squeezed attention module proposed in this paper can integrate multi-scale spatial information and cross-channel attention into blocks of each feature group. Therefore, the model implemented in this paper can obtain better information interaction between local and global channel attention.

3.3 Occlusion perception (Op)Module

Distinguishing from previous top-down instance segmentation methods, this paper proposes a two-layer decoupling model based on Occlusion perception models the region of interest in an image as two overlapping layers, with the upper layer detecting the occluded target and the bottom layer reasoning about the occluded tar-get. The explicit modeling method of the dual-layer structure separates the boundary between the occluder object and the occludee object, and achieves the consideration of the interaction between the occluder object and the target through the prediction of the occludee object and the boundary, thus improving the processing capability of the image instance segmentation model for complex occluded objects, as shown in figure 4. The top GCN layer detects the occluder object and the bottom GCN layer infers the instance of the occluded garbage.

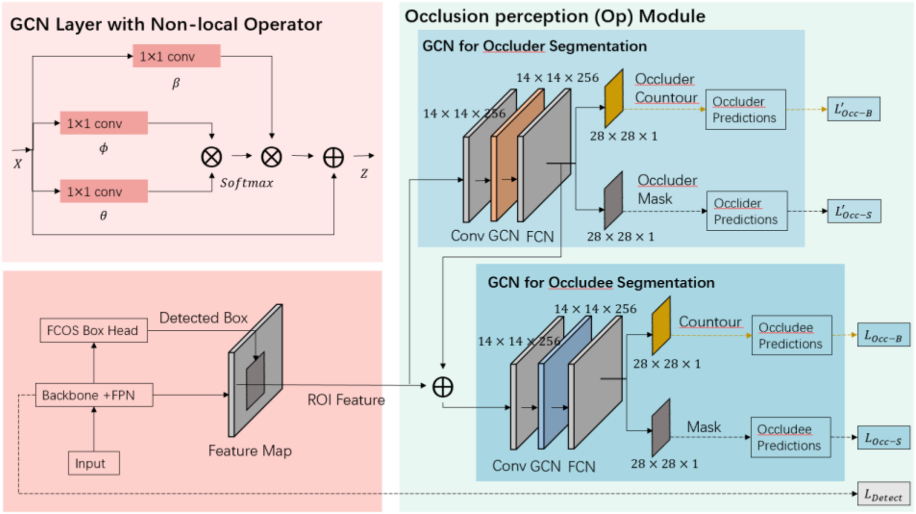

Fig. 4. Double decoupled structure.

Figure 5 shows a schematic diagram of the two-layer decoupling structure, including the top layer as well as the bottom layer. The overlapping part of the two is the invisible region of the occludee object, which is displayed and modeled by the two-layer decoupling model. The first layer, GCN, provides a large amount of occlusion information such as the shape and position of the occludee object, and guides the instance segmentation process of the occluded image.

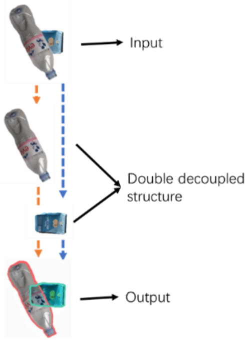

Fig. 5. Invisible Occluded Region.

The input x denotes the CNN feature after ROI extraction. Conv has 3 × 3 core convolution layer, FC is the full connection layer, and SAM is the spatial attention module. Bt and Mt refer to the box and mask head at t-th stage. Different from the previous occlusion perception mask head, it regress both modal and amodal masks from the occludee. Our module has a double-layer GCN structure, and takes into account the ROI of the same interaction between the top "occluder" and the bottom "occludee". The occlusion occludee branch explicitly models occluded objects by performing joint masks and contour prediction, and extracts basic occlusion information for the second layer to segment the target object ("occludee").

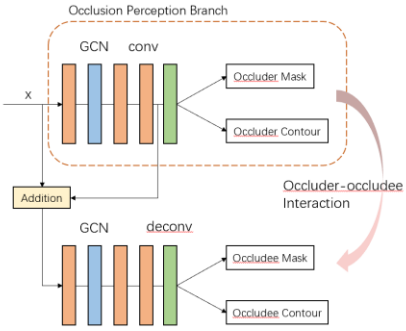

Fig. 6. Occlusion Perception Branch.

The system consists of a garbage detection part and a garbage segmentation part. Where the segmentation network can be represented as

$$Z = \sigma(AXW_g) + X, \tag{9}$$

where $X \in R^{N\times K}$ is the input feature, N=H×W is the number of pixel grids in the RoI region, K is the feature dimension of each node, $A \in R^{N\times N}$ is the adjacency matrix used to define the adjacency of graph nodes by feature similarity, and $W_g \in R^{K\times K'}$ is the learnable weight matrix of the output transformation, where in the case of this paper K′= K. The output feature Z∈$R^{N\times K'}$ consists of the node features that are updated by propagating them through the node features updated by global information propagation within the whole layer, which are obtained after a nonlinear function σ（·）including layer normalization and ReLU function. In this paper, a residual connection is added after the GCN layer.

To construct the adjacency matrix A, the pairwise similarity between every two graph nodes x_i, x_j is defined in this paper by dot product similarity as

$$A_{ij} = softmax\left(F\left(x_i, x_j\right)\right); \tag{10}$$

$$F\left(x_i, x_j\right) = \theta(x_i)^T \phi\left(x_j\right)^T, \tag{11}$$

where θ and φ are two trainable transform functions implemented by a 1 × 1 convolution of the nonlocal operator part, making the high confidence edges between two nodes correspond to greater feature similarity.

The two-layer decoupling structure will input the extracted ROI features X_{roi} to the first GCN layer to get the updated features Z^0 and derive the contours and masks of the occluder objects. The updated feature Z^0 is then added to the ROI feature as the input to the second GCN layer ($X_f = X_{roi} + Z^0$). The second GCN layer will further derive the contours and masks of the occludee objects.

For the example of occlusion prediction in the presence of occlusion, using the occlusion-aware model would encode the occluded and occluder layers using two separate occlusion prediction layers, and later fuse the results of the two layers to obtain the final result.

4 Experiments

4.1 Datasets

In this paper, we use the publicly available Taco garbage image dataset [11] and Huawei Cloud's household garbage dataset [12] to build the garbage image dataset used in this paper for model training and testing. The constructed dataset includes different lighting conditions (e.g., strong, weak, nighttime, etc.) and different backgrounds (e.g., beach, road, grassland, etc.). The collected images were labeled using Labelme, and the labeled images were classified into 60 categories: Battery, Food Can, Paper cup, etc.

4.2 Main Results

Quantitative evaluation. In this paper, the proposed algorithm is compared with seven typical instance segmentation detection methods, including YOLACT[13], ContrastMask[14], Mask R-CNN[15], PANet[16], SOLO[17], Cascade R-CNN[18], and HTC[19].

Table 1. Quantitative evaluation

Method	mAP	AP0.5	AP0.75
YOLACT	51.3	70.7	49.8
ContrastMask	54.6	78.6	54.1
Mask R-CNN	55.2	80.2	57.6
PANet	57.3	81.4	59.4
SOLO	58.0	81.7	60.8
Cascade R-CNN	58.4	82.1	61.3
HTC	58.5	82.3	61.5
Op-PSA	59.0	83.1	62.9

As can be seen from Table 1, the Op-PSA model studied in this paper achieves the highest detection accuracy with 83.1% of garbage detection accuracy. The Op-PSA model can obtain higher accuracy and recall compared with the comparison method, which is attributed to the fact that the Op-PSA model can more accurately separate the masked objects from the masked garbage and solve the low detection rate in the case of masking in the data set. The model can solve the occlusion problem in the accuracy of the pyramidal squeezed attention model and improve the model's resistance to occlusion.

Qualitative Evaluation. In order to compare subjectively, this article extracted and tested garbage images under occlusion from the dataset, and arranged the garbage images detected by various algorithms as follows. It can be seen that the Op-PSA model has better detection performance for garbage under occlusion, and has higher detection accuracy in terms of human perception.

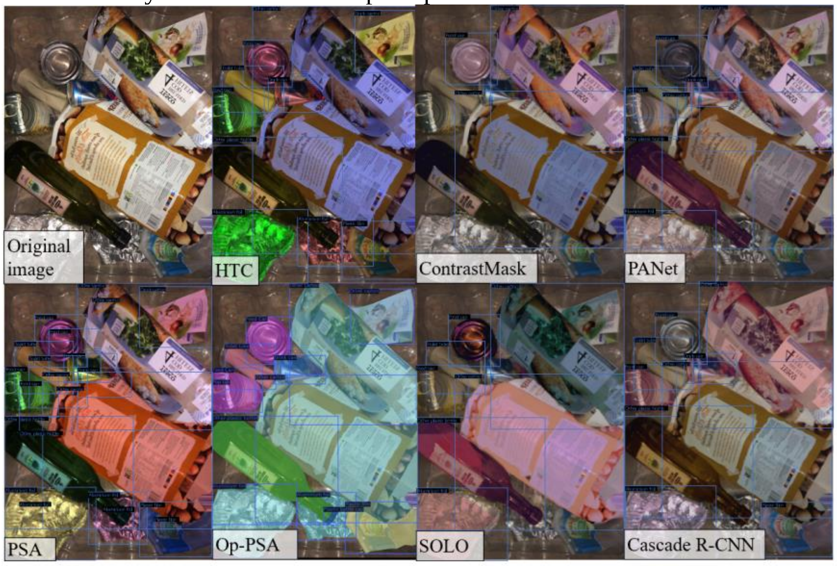

Fig. 7. A large number of occluded garbage maps.

From the above, the Op-PSA model can give better detection results for the dataset used in this paper, even in the case of small-scale occlusion. And by visualizing the front and back layers separately and modeling the boundary and mask of the occluded and masked objects, the detection of occlusion situations can be better grasped, with better recognition of the occluded garbage.

4.3 Ablation Study

Effectiveness of the Pyramid Squeezed Attention (PSA) Module. In this paper, an ablation study is conducted to evaluate the impact of location when integrating the pyramid squeeze attention module into an existing architecture. In addition to the proposed design, three variants are considered in this paper: (1) a front position, where the PSA module is moved before the residual unit; (2) a back position, where the PSA module is after the residual unit; and (3) a parallel position, where the PSA module is placed on a sign connection parallel to the residual unit. These variants are shown in Figure 8, and the detection results of each variant are shown in Table 2.

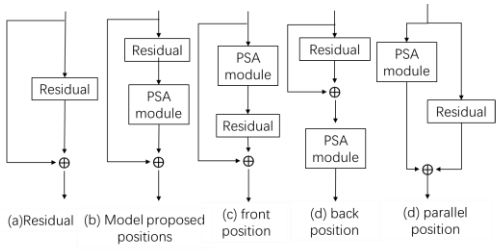

Fig. 8. Comparison diagram of ResNet and module position.

Table 2. Ablation Experiments with Adaptive PSA Module

PSA module	mAP	AP0.5	AP0.75
Front position	58.5	82.3	61.5
Back position	57.2	81.2	59.8
Parallel position	58.7	82.5	62.0
Proposed position	58.7	82.8	62.0

From the above table, it can be observed that the modules for the front module, the parallel module, and the proposed position are detected well, while the model using the back module leads to performance degradation. This experiment shows that the adaptive pyramid squeezes attention module produces performance improvements that are stable for each location as long as they are applied before branch aggregation.

Effectiveness of the two GCN layers. A second GCN layer is added to the model and the final occlusion prediction of the effect of the second GCN on detecting the occlusion contours is guided by the output of the first GCN. That is, in the form of cascade optimization, the second level of optimization is performed on top of the first level after the prediction of the occlusion layer is completed.

Table 3. Ablation experiments with two GCN layers

First GCN layer guidance	Contour	Mask	mAP	AP0.5	AP0.75
—	—	√	52.8	78.4	57.1
√	—	√	56.3	80.3	60.4
√	√	√	59.0	83.1	62.9

As can be seen from Table 3, the model is guided by the first GCN output for the occlusion prediction of the second GCN layer to obtain a more accurate detection structure. Taking AP0.5 as an example, for the original model, the model with the two-layer decoupled structure is more accurate for garbage detection under occlusion, with a 1.7% increase in accuracy. With the addition of Contour, the model achieves the optimal detection result of 83.1%, which is 1.7% and 2.6% better than the model without the first GCN layer and the model without Contour, respectively. It can be seen that the two-layer decoupled model can indeed improve the accuracy rate of the improved HTC model for garbage detection by improving the garbage detection results in the case of occlusion, proving that this model has high robustness and accuracy.

5 Conclusion

In general, this paper proposes an attention model-based garbage instance segmentation detection method for obscured garbage based on the problems of difficult feature extraction, low feature information utilization, and more occlusion cases in the instance segmentation garbage detection method. The experimental results show that the improved network model can better extract the features of garbage information and can effectively improve the accuracy of garbage detection.

References

1. Bashkirova D, Abdelfattah M, Zhu Z : Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In: Rama Chellappa(John Hopkins Univ.) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, June 19-20,21147-21157. IEEE, New York City (2022)
2. Zhang C, Liu X:Feature extraction of ancient Chinese characters based on deep convolution neural network and big data analysis. Computational Intelligence and Neuroscience-31(24),249-256(2021)
3. Guo C, Fan B, Zhang Q: Augfpn: Improving multi-scale feature learning for object detection. In: Terry Boult, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, Jun 13- 19,12595-12604. IEEE, New York City (2020)

4. Fulton M, Hong J, Islam M J: Robotic detection of marine litter using deep visual detection models. In: Yoshua Bengio. 2019 international conference on robotics and automation (ICRA),Canada, May 20-24, 5752-5758. IEEE, New York City (2019)
5. R.Xu, J.An, L. Su: Banknotes serial number coding recognition. In: Dr. Roger Barga (Amazon.com),2019 IEEE International Conference on Big Data (Big Data), Los Angeles,December 9-12, 5101-5107. IEEE, New York City (2019)
6. Wang T, Cai Y, Liang L:A multi-level approach to waste object segmentation. Sensors-20(14),3816-3817(2020)
7. Rajaei K, Mohsenzadeh Y, Ebrahimpour R:Beyond core object recognition: Recurrent processes account for object recognition under occlusion. PLoS computational biology-15(5),1007-1008(2019)
8. TIAN Y L, LUO P, WANG X G: Deep learning strong parts for pedestrian detection. In: Sawada T ,Proceedings of the 2015IEEE International Conference on Computer Vision, San-tiago, Dec 11-18, 1904-1912. IEEE Computer Society, Washington(2015)
9. CHU X G, ZHENG A L, ZHANG X Y: Detection incrowded scenes: one proposal, multiple predictions. In: Terry Boult, Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13- 19, 12211-12220. IEEE, Piscataway(2020)
10. WANG X L, XIAO T T, JIANG Y N: Repulsion loss: detecting pedestrians in a crowd. In: Michael Brown ,Proceedings of the IEEE conference on computer vision and pattern recognition, Salt lake city, Jun 19-21, 7774-7783. IEEE Computer Society, Washington(2018)
11. Wang T, Cai Y, Liang L: A multi-level approach to waste object segmentation. Sensors, 20(14),3816(2020)
12. 2019 Huawei Cloud AI Competition. Garbage sorting data, https://aistudio.baidu.com/aistudio/datasetdetail/ 16284.html,last accessed 2019/06/20
13. Bolya D, Zhou C, Xiao F: Yolact: Real-time instance segmentation. In: Larry Davis, Proceedings of the IEEE/CVF international conference on computer vision. CA, June16-209157-9166. IEEE, New York City (2019)
14. Wang X, Zhao K, Zhang R: Contrastmask: Contrastive learning to segment every thing. In: Rama Chellappa(John Hopkins Univ.) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, June 19-20, 11604-11613. IEEE, New York City (2022)
15. He K, Gkioxari G, Dollár P: Mask r-cnn. In: Katsushi Ikeuchi(Microsoft Research Asia)Proceedings of the IEEE international conference on computer vision, Venice, October 22-29, 2961-2969. IEEE, New York City (2017)
16. Wang K, Liew J H, Zou Y: Panet: Few-shot image semantic segmentation with prototype alignment. In: Larry Davis ,proceedings of the IEEE/CVF international conference on computer vision, CA, June16-20, 9197-9206. IEEE, New York City (2019)
17. Wang X, Kong T, Shen C: Solo: Segmenting objects by locations. In: Vittorio Ferrari ,Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 649-665. Springer International Publishing, (2020)
18. Cai Z, Vasconcelos N: Cascade r-cnn: Delving into high quality object detection. In: Michael Brown ,Proceedings of the IEEE conference on computer vision and pattern recognition, Salt lake city, Jun 19-21, 6154-6162. IEEE, New York City (2018)
19. Chen K, Pang J, Wang J: Hybrid task cascade for instance segmentation. In: Larry Davis ,Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, CA, June16-20, 4974-4983. IEEE, New York City (2019)