Object detection in seriously degraded images with unbalanced training samples*

LIU Sheng (刘盛), SHEN Jiayu (沈家瑜), and HUANG Shengyue (黄圣跃)**

Collage of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

(Received 12 January 2021; Revised 25 February 2021)

Uncertain environments, especially uneven lighting and shadows, can degrade an image, which causes a great negative impact on object detection. Moreover, unbalanced training samples can cause overfitting problem. Since available data that is collected at night is much rarer than that collected in the daytime, the nighttime detection effect will be relatively poor. In this paper, we propose a novel data augmentation method named Mask Augmentation, which reduces the brightness and contrast of objects, and also weakens the edge of objects to simulate the degraded scene. In addition, we propose a new architecture, by adding a classification loss branch and a feature extraction module named Multi-Feature Attention Module, which combines the attention mechanism and feature fusion on the basis of Darknet-53. This architecture makes the features extracted in daytime and nighttime images distinguishable. We also increase the loss weight of nighttime images during the training process. We achieved 78.68% mAP on nighttime detection and 73.14% mAP on daytime detection. Compared with other models, our method greatly improves the accuracy of nighttime detection, and also performs satisfactorily on daytime detection. We deployed our model on an intelligent garbage collection robot for real-time detection, which implements automatic picking at night and assists cleaning staff during the day.

Document code: A **Article ID:** 1673-1905(2021)09-0564-8

DOI https://doi.org/10.1007/s11801-021-1005-6

Object detection has always been a hot topic in the field of computer vision. Two-stage models, such as Faster Region-Based Convolutional Neural Network (Faster R-CNN)[1] first generate proposals through Region Proposal Network (RPN), and then output more accurate results after classification and regression. Because of these two steps, the detection accuracy is high; however, the speed is slow. Later the multi-stage model Cascade R-CNN[2] trains multiple cascaded detectors, making the accuracy higher. One-stage models run much faster with relatively high accuracy. YOLO v3[3] uses anchor mechanism and feature fusion, and has the ability to detect overlapping and multi-scale objects. FCOS[4] regresses each point on the feature map and realizes multi-scale prediction based on Feature Pyramid Network (FPN). FoveaBox[5] is similar to FCOS, but it allocates boxes of different sizes to feature maps of different levels according to area, which can increase robustness. RepPoints[6] makes flexible use of deformable convolutions and locates some noteworthy points of objects to generate prediction boxes. With time going by, more and more high-accuracy real-time detection models are proposed, which bring very powerful effects and penetrate the industry gradually.

However, environmental factors cause image degradation, which can worsen the object detection effect. Objects and their surroundings may have similar colors or textures, resulting in blurred edges. Additionally, due to strong light, the surfaces of objects reflect light and cause a lot of color information loss. Moreover, in shadows and dark environments, objects lose brightness and contour information. If important original information is disturbed or even lost, the features extracted by the convolutional neural network will be unreliable, which can affect the detection results.

In addition, unbalanced training samples will lead to overfitting of detection models. Due to the high cost of partial data collection, training samples are not evenly distributed. If a class of data has more samples, it will occupy a greater weight in the overall loss. Therefore, during the gradient descent, detection models will overfit the data which has a larger number and perform better on this data.

Garbage detection has a problem of image degradation and sample imbalance. Unfortunately, in our dataset, the amount of nighttime data is only 1/20 of the amount of daytime data, and the nighttime data are highly degraded. Because of the dark environment at night, the edges and colors of garbage are almost indistinguishable. Daytime

* This work has been supported by the National Key Research and Development Program of China (No.2018YFB1305200), and the Science Technology Department of Zhejiang Province (No.LGG19F020010).

** E-mail: edliu@zjut.edu.cn

data are also degraded due to the complex environment, strong light, shade of the tree, or camera shake.

We designed an intelligent garbage collection robot equipped with a garbage detection model. In order to make the robot work at night and assist the cleaning staff during the day, we need to improve the accuracy of nighttime garbage detection as much as possible, while ensuring that the accuracy of daytime garbage detection is not decreased.

We proposed a solution and it produced good results. As shown in Fig.1, we compared the proposed method with YOLO v3 on the public dataset TACO[7] and our own dataset. In Fig.1(a), our method detects the other plastic product which is disturbed by a complex environment. In Fig.1(b), our method does not treat the light spot as a bottle. In Fig.1(c), even if the nighttime image has a lot of noise, our method can still detect all objects. In Fig.1(d), even if the environment is very dark, our method can still detect most objects. Our method effectively improves the nighttime detection accuracy with unbalanced training samples, and enhances the robustness to degraded images. The main contributions of our work can be summarized as follows:

(1) We design a data augmentation method named Mask Augmentation, which covers the image with a specific mask to form shadows and night effects. Mask Augmentation can alleviate the problem of degradation, increase the number of training samples to solve overfitting, and effectively improve the detection accuracy of nighttime data.

(2) We propose a new module named Multi-Feature Attention Module, which combines feature fusion and attention mechanism. Multi-Feature Attention Module can improve daytime detection accuracy while maintaining high nighttime detection accuracy.

(3) We add a classification loss branch, which enables the backbone to learn the different features of daytime and nighttime images, and then improve the nighttime detection accuracy.

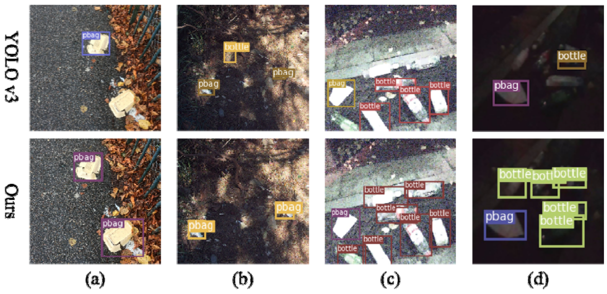


Fig.1 Comparison of our method and YOLO v3: (a) A complex environment situation from the TACO dataset; (b) A situation where light spots are formed under strong light from the TACO dataset; (c) The nighttime images with a lot of noise taken by a starlight camera from our dataset; (d) The nighttime images taken by an ordinary camera from our dataset which is very dark

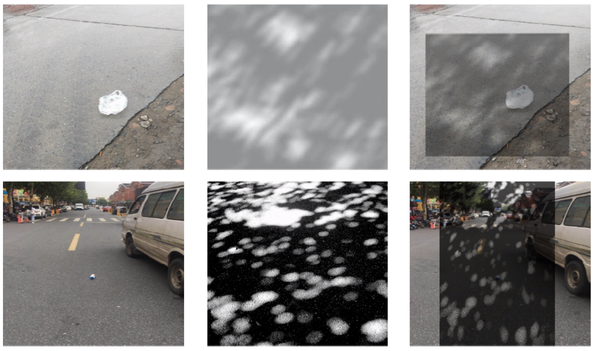
Object detection has made tremendous progress in recent years. From the earliest artificial feature extraction to the current automatic feature extraction through convolutional neural network, the detection accuracy has been significantly improved. The traditional method uses the SIFT[8] algorithm to extract feature vectors after obtaining candidate regions, and then passes these feature vectors through the SVM[9] classifier. Later, deep learning methods become popular. The two-stage detection model R-CNN[10] generates candidate regions and then uses convolutional neural networks for classification and regression. Fast R-CNN[11] solves the problem of repeated calculations. Faster R-CNN generates candidate regions through convolutional neural networks to increase speed. Although two-stage detection models have high accuracy, they run slowly and cannot achieve real-time performance. YOLO is a real-time object detection model which has many variations. YOLO v1[12] uses Darknet as the backbone to improve detection speed. Each channel of the output tensor represents the position, size, object class, confidence score, and other information of the corresponding detection frame. The confidence score represents the probability that each box contains an entity. Finally, the detection object is screened by non-maximum suppression. YOLO v2[13] adds the anchor mechanism, which makes convergence easier and solves the overlap problem to a certain extent. YOLO v3 adds feature fusion, which combines concrete and abstract layers, and outputs tensors of three scales, which effectively solves the problem of small object detection. After that, some models, such as RetinaNet[14], propose new loss function, and some models, such as CenterNet[15], propose new key point representation methods. In the most recent year, Camouflaged Object Detection[16] solves the problem of detecting objects embedded in the environment, Few-Shot Object Detection[17] solves the problem of model training through a small amount of annotated data, and D2Det[18] proposes a method to improve positioning accuracy and classification accuracy.

However, these models do not solve the problem of object detection on degraded images nor the problem of unbalanced training samples. Data augmentation such as flipping, affine, and scaling solves the overfitting problem, but object detection in degraded images is still a big problem. Adjusting the loss weight ratio and changing the sampling distribution are effective ways to alleviate the problem of unbalanced training samples, but there are still many things to do to improve in overall accuracy.

Therefore, we designed a data augmentation method named Mask Augmentation and a feature extraction module named Multi-Feature Attention Module to solve these problems. It turns out that despite the small sample size of nighttime images, our method can still have considerable accuracy in nighttime detection, and it also performs well in daytime detection. At the same time, our method can effectively solve the problem of object detection in degraded images.

Data augmentation is a way to increase sample diversity and reduce overfitting, but the basic augmentation operation cannot solve the problem of object detection in degraded images, so we designed an augmentation method for object detection on degraded images, named Mask Augmentation.

The mask we designed can correspond to various scenes. We added random light spots on a black background, to simulate the shadow of trees under strong light and the interference of street light at night as much as possible. We randomly generated many such masks. These masks can change the color distribution of data, and they reduce the brightness of garbage objects, so the edges of garbage objects have a certain blur. Moreover, these masks can make the model better accommodate degraded features during training. Some masks are even completely black, while some masks are almost full of light points. These are considered according to different situations. For example, sometimes a nighttime image has no light source, and the scene is black. Sometimes the garbage just appears beneath a street light, and a large area in the middle is particularly bright. Fig.2(b) shows the designed masks where the color depth as well as the area of light spots are different. These are randomly generated in order to increase the diversity of augmented samples and adapt to more degradation situations.



(a) Original (b) Mask (c)Augmented

Fig.2 The result of adding a mask to the original image: (a) Original images; (b) Designed masks; (c) Images after adding the mask

The mask can cover all detection objects. We added the mask to an image in a random way. Because the ambient light intensity is uncertain, the shadow color depth is different. The darkness is different at night due to different times and effects of the ground material, so the transparency of the mask is set to a random value. The random transparency makes the color depth of the mask different, which can be better generalized into various degradation scene. We also set a ratio of the mask area and aspect to random values, because from different shooting angles, the area and aspect occupied by the shade of tree change. In addition, the distance from a street light at night also causes the size of the dark area to be different. Fig.2(c) shows the images after adding the mask. The area ratio and aspect ratio occupied by the mask in the image are uncertain, also for diversification. Under the cover of these two masks, the edges and colors of objects in these two images are blurred to varying degrees, making the trained model more robust to degraded images.

We used Darknet-53 as backbone and proposed a new network architecture. As shown in Fig.3, the features obtained by the image after darknet-53 are input to the classifier and Multi-Feature Attention Module to obtain four outputs, and the corresponding losses are optimized. y1, y2, y3 are output tensors at different scales, from which the information of the detected object can be extracted. There are many basic modules in backbone, such as DBL and resn. These modules are very simple, DBL is only composed of a convolutional layer, a batch normalization[19] layer and Leaky Relu[20] activation function, the combination of these simple ingredients can quickly extract features. Then DBL is followed by a large number of residual layers[21], which are proved to make the network deeper and easier to train, and let the training loss converge lower. When connecting several residual layers, the tensor needs to be downsampled once, which can abstract the features and enable the model to learn deeper information. For convenience of presentation, we combined downsampling with different numbers of residual layers to form a resn block, where *n* represents the number of specific residual layers included. For the last resn block, the feature tensor generated by it passes through the classifier we designed. At the same time, the output from each of the last three resn blocks are input to the Multi-Feature Attention Module together to produce three scale output tensors, which have different scales and contain information about the detected objects. We can extract features on the three tensors and use Non-Maximum Suppression (NMS)[22] to obtain the detection result, which is consistent with YOLO v3.

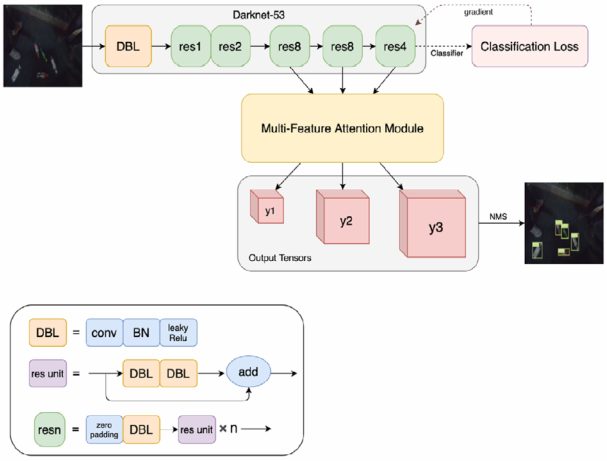


Fig.3 Network architecture

Multi-Feature Attention Module is our mainly improved structure. We proposed an innovative method to introduce attention mechanism. Different feature layers contain different information. The fusion of shallow specific information and deep abstract information can obtain more accurate attention, which is not just limited to channels and spaces, able to each point. As shown in Fig.4, we extracted the last three resn blocks of the Darknet-53 output. The features extracted from the last layer are more abstract, so they can describe the semantics more easily. However, due to downsampling, the receptive field is also very large, these features are often used to detect large objects. Although the features extracted from the shallow layer are more specific, they are not lost too much original information by convolution, and have a relatively large space, making it easier to detect small objects. In order to allow the generated attention to take into account both abstract and concrete information, we upsampled the features of each resn block after a series of convolutions, then fused the features of the previous resn block. This operation is carried out twice, so that the last three resn blocks of Darknet can be combined in a certain order, which can obtain strong semantics without losing a part of important original information. This is beneficial for obtaining more powerful attention. Considering the detection efficiency, we set the number of convolutions to 2. Because the attention mask we designed is point-level, it needs to be downsampled to be consistent with the shape of the deep feature. After the sigmoid function, the value of the attention mask is normalized in range of 0 to 1. This can evaluate the importance of each point of deep feature. Next, we fused the features that have been fused in the first round. The difference is that we added attention to the deep feature, and the number of convolutions has been changed from 2 to 5. We only added attention to the feature generated by the last resn block, because the attention can affect the results of three outputs through feature fusion. In Fig.4, the feature fusion of the first stage produces an attention mask and new features for the second stage, then the attention mask is multiplied with the new deep features, and the fusion again outputs three tensors.

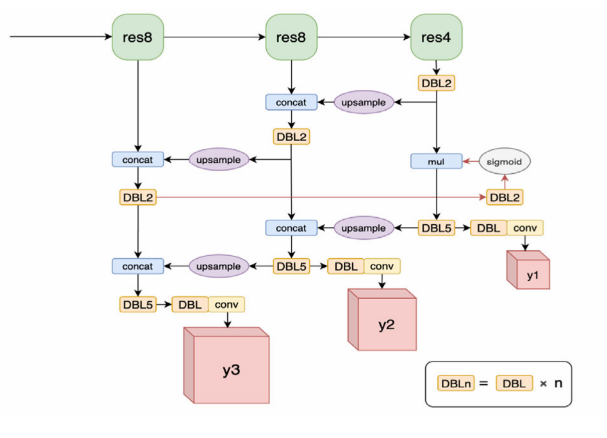


Fig.4 Multi-feature attention module

The loss branch is generated by a classifier. This classifier divides the images into two categories: daytime and nighttime images. As shown in Fig.5, the structure is simple. The last output feature of Darknet-53 is passed through a global average pooling[23] layer to obtain a long strip feature. The value output by the final fully-connected layer represents the probability of a nighttime image, where 1 means that the model completely thinks that this image is taken at night, and 0 just the opposite. Since there are only two classes, we used the binary cross-entropy loss as the function.

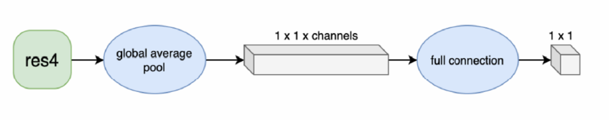


Fig.5 The classifier structure, where the features generated by Darknet output a binary value through a series of pooled and fully connected, representing the confidence of the nighttime image

This classifier is not used to classify the detected object but to separate features. By gradient descent of the classification loss, the previous layers can learn the different features of daytime and nighttime images. Therefore, the model can more clearly distinguish the different characteristics of the same entity during the day and night. The same garbage object looks different during the day and night. It is brighter during the day and even has strong reflections. At night, it is overall dark and easily to merge with the surrounding environment. The addition of classification loss makes the features have a certain degree of separation, which helps to improve the accuracy of nighttime detection for garbage.

Due to the huge difference in the number of daytime and nighttime images, if the model is trained directly, it needs to pay more attention to the accuracy of the daytime detection, which accounts for almost all of the total loss, so the performance of nighttime detection is not good. In order to make the model pay attention to the accuracy of the nighttime detection during training, we increased the loss weight of nighttime image.

$$Loss_x = \sum_{output} loss(output, target),$$
$$Loss_{\mathrm{all}} = \alpha \sum_{x \in nighttime} Loss_x + \sum_{x \in daytime} Loss_x . \qquad (1)$$

As shown in Eq.(1), $Loss_x$ represents the loss between the output of current model and the ideal value when image x is used as input. $Loss_{\mathrm{all}}$ represents the loss of all images, of which the loss of nighttime images has a weight coefficient α, we can adjust this value to change the position of nighttime images in training. If α is set too small, the model will still be biased towards the daytime data. On the contrary, if the setting is too large, the model will overfit the nighttime data, and the performance on the daytime data will be very poor. By comparison, we set α to 16. Although a hyperparameter is introduced, it can slow down the problem of unbalanced training samples and improve the accuracy of nighttime garbage detection.

We built a dataset for which the training set contains 8 000 daytime images and 400 nighttime images, and the test set contains 500 daytime images and 500 nighttime images. As shown in Fig.6, our dataset contains various scenarios. We asked environmental protection departments to provide data in multiple environments such as riverside, roadside, and grass. We also increased degraded images as much as possible, including under strong light, under the shadow of trees, and in dark corners. For nighttime images, we also almost cover a variety of scenes, including under street lights and on completely dark roads. We also use multiple different cameras for shooting, including ordinary cameras and starlight cameras, and the images they take have different degrees of degradation. All our nighttime images are degraded. We first label all images in two categories: daytime image and nighttime image, and then draw label boxes for all the garbage and divide them into five categories: bottle, cup, pbag (including all non-box garbage related to plastic and paper), box, and butt. At the same time, we make corresponding annotations on the public dataset TACO, and delete the indoor images to facilitate testing. During the model training process, the resolution of all images is compressed to 416×416, and a certain scale and flip transformation are performed.

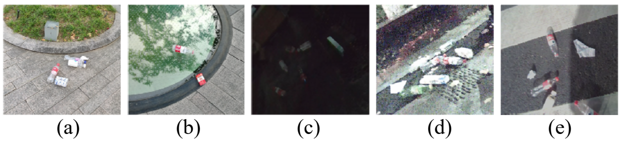


Fig.6 Five images are from our dataset: (a) A normal daytime image; (b) Image degraded due to specular reflection; (c) Image taken with ordinary camera at night; (d) image taken with starlight camera at night; (e) image taken under light at night

We chose Mean Average Precision (mAP) as the evaluation metric, which is a common metric in the field of object detection. This value is related to the Precise and Recall of the detected object. Precise and Recall are in a contradictory state, we can draw a PR curve based on this, AP is the area covered by the PR curve, and the average result for all classes is mAP. For data in COCO format[24], we took AP50 as mAP.

We chose different loss functions for different output features. For the boundary information such as coordinate points offset, length ratio and width ratio of the detection box relative to the anchor, we used the mean square error loss, which can more directly return the estimated value to the value close to ground truth, reduce manual labeling errors and interference of degraded edges. For the confidence, the classification of garbage objects, and the classification of images, we used the binary cross-entropy loss function, which can make the probability distribution close and help training. When calculating the total loss, the weights of Classifier Loss is 0.5, and the others are 1.

Our method is not very complicated, but it has an obvious effect in the experiments. Although the number of nighttime images is small, it greatly improves the accuracy of nighttime garbage detection. At the same time, it also performs well in the daytime detection. Our proposed method can also effectively solve the problem of object detection in degraded images.

Tab.1 The results of ablation experiment

Method	Dataset	N-mAP (%)	D-mAP (%)	All-mAP (%)
(1) YOLO v3 (benchmark)	N	76.67	16.50	46.58
(2) YOLO v3 (benchmark)	N+D	67.12	69.44	68.28
(3) FEY+MA	N+D	73.69	**77.67**	75.68
(4) FEY+MA+ LWRA	N+D	77.53	68.18	72.85
(5) FEY+MA+ LWRA+CL	N+D	**79.58**	66.51	73.04
(6)MFAM+MA+ LWRA+CL	N+D	78.68	73.14	**75.91**

N-mAP: The mAP of nighttime data in the test set; D-mAP: The mAP of daytime data in the test set; All-mAP: The mAP of all data in the test set; N: Nighttime Data; D: Daytime Data; FEY: Feature Extractor of YOLO v3; MA: Mask Augmentation; LWRA: Loss Weight Ratio Adjustment; CL: Classification Loss; MFAM: Multi-Feature Attention Module

Tab.2 The comparison results with other models

Model	Backbone	N-mAP (%)	D-mAP (%)	All-mAP (%)
Faster R-CNN + FPN	Resnet-50-FPN	73.8	74.3	74.0
YOLO v3	Darknet-53	67.12	69.44	68.28
FoveaBox	Resnet-50-FPN	63.5	65.4	64.4
FCOS	Resnet-50-FPN	67.1	76.7	71.9
RepPoints	Resnet-50-FPN	64.2	69.7	66.9
YOLO v4	CSPDarknet-53	56.36	88.68	72.52
(1) FEY+MA	Darknet-53	73.69	77.67	75.68
(2)FAM+MA+ LWRA+CL	Darknet-53	**78.68**	73.14	**75.91**

We performed many sets of experiments to verify the correctness of our method. Before experimenting, we first controlled the tunable parameters, such as the optimizer and learning rate, to remain unchanged, and used the pre-trained model for Darknet-53 on ImageNet[25] to increase the convergence speed. Most experiments reach the convergence state after iterating 100 epochs, and we took the best performing result as the experimental result. For each method, we calculated the mAP values of nighttime data, daytime data and all data in the test set. As shown in Tab. 1, all proposed methods receive different effects. If the various methods are superimposed, more satisfactory results can be produced. For a more intuitive comparison, we also separately trained nighttime data on YOLO v3.

Benchmark Tab.1(1) shows that if the night data is used for training alone, the model is easy to overfit because of the small number. Although the performance of the nighttime data detection is good, the mAP can reach 76.67%, but the performance in the daytime data is particularly poor, the mAP is only 16.50%. Tab.1(2) shows that if we train the daytime data and the nighttime data together, because there are many daytime data, the model is more likely to pay attention to the accuracy of the daytime detection, the performance on the nighttime data is reduced, and the mAP is lower than Tab.1(1) 9.55%. At the same time, after the model sees a lot of daytime data, the ability to detect daytime data is greatly improved, mAP reaches 69.44%.

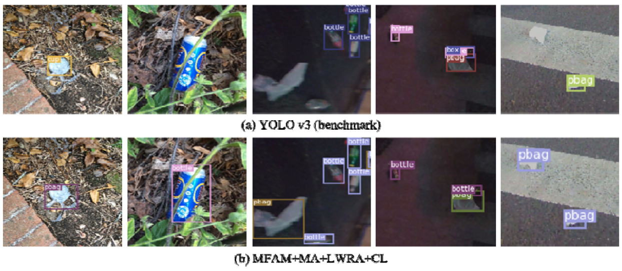
Add Mask Augmentation Tab.1(3) shows that the detection performance is greatly improved after the addition of Mask Augmentation. Compared with Tab.1(2), the mAP of the nighttime data is increased by 6.57%, and the mAP of the daytime data is increased by 8.23%. The important reason for this result is the augmentation method we proposed can improve the robustness to degraded images, and the detection accuracy of objects in the corner or shadow is greatly improved. Since the nighttime image is degraded, Mask Augmentation also has a very obvious effect. This augmentation method darkens the local area on the image to a certain extent, and increases the sample to avoid overfitting.

Add Loss Ratio Adjustment Tab.1(4) shows that the adjustment of loss weight ratio can make loss weight of nighttime data greater, the mAP of nighttime data is greatly improved. The mAP is increased by 3.84% compared to Tab.1(3) and 0.86% compared with Tab.1(1). Because the model pays more attention to the performance on nighttime images, the nighttime detection effect is better than Tab.1(4). And because the addition of daytime data reduces overfitting, the nighttime detection effect is better than Tab.1(1). But the performance on the daytime data is greatly reduced, mAP is only 68.18%.

Add Classification Loss As shown in Tab.1(5), adding a new Classification Loss branch can implicitly separate nighttime features, thereby increasing the mAP of nighttime data to 79.58% again. However, the mAP of the daytime data is not satisfactory, only 66.51%.

Add Multi-Feature Attention Module Tab.1(6) shows that after adding the Multi-Feature Attention Module, the mAP of the nighttime data is 0.9% lower than Tab.1(5), but the mAP of daytime data is 6.63% higher than Tab.1(5). Although the accuracy of nighttime detection is not much different, the accuracy of daytime detection has a considerable result. This is because the module we designed can easily combine the information of specific layers and abstract layers to generate different attention to the daytime and nighttime data. Compared with the benchmark, the mAP of the nighttime data is 2.01% higher than Tab.1(1) and 11.56% higher than Tab.1(2), the mAP of daytime data is 56.64% higher than Tab.1(1) and 3.7% higher than Tab.1(2). The mAP of all data is the highest.

As shown in Fig.7, we visualized the experimental results and compared the detection results of Tab.1(2) and (6) on select images. If the benchmark method is used, in the first image, when the environmental interference is relatively large, the plastic cap is misclassified. In the second image, the bottle is not detected. In the third image, some garbage is not detected. In the fourth image, there is a false positive detection result. In the fifth image, the paper ball is not detected due to the sidewalk interference. Our method performs better in these situations. In Fig.7, some of garbage in these images is disturbed by the environment, and some is seriously degraded at night. The method we proposed has a significant improvement in detection compared with the benchmark.



(a) YOLO v3 (benchmark)

(b) MFAM+MA+LWRA+CL

Fig.7 Visualization of detection results, where the first two columns are taken from the TACO dataset, and the last three columns are taken from the self-built dataset

We compared our method with other object detection models. The two-stage models, such as Faster R-CNN, have high detection accuracy, but the speed is low. One-stage models such as FCOS and FoveaBox can balance the accuracy and speed, and achieve good real-time detection results. As shown in Tab.2, we compared our method with other models. Except for YOLO v4[26], our method(1) has the highest daytime detection accuracy. Compared with all other models, method(2) has the highest nighttime detection accuracy and considerable daytime detection accuracy. The state-of-the-art model YOLO v4 performs very well in daytime detection, but performs poorly in nighttime detection. The nighttime accuracy and overall accuracy of method(2) is higher than that of YOLO v4. Although Faster R-CNN has very impressive daytime and nighttime detection accuracy, it runs very slowly and cannot perform real-time detection. Therefore, our method has great advantages and can be used in intelligent real-time detection machines.

We applied the proposed method to our intelligent garbage collection robot. Since our dataset contains a large amount of daytime data and a small amount of nighttime data, there is a great imbalance. Our method can improve the nighttime detection accuracy as much as possible and ensure that it also has no negative effects on daytime detection. To a certain extent, we also solved the problem of object detection in degraded images. Supported by this method, our intelligent garbage collection robot can work independently at night and assist environmental protection staff during the day. Because the robot travels along the roadside, it is found that the garbage objects are often covered by the shade of the trees during the actual real-time detection, which causes degraded edges and increases the difficulty of detection. Our proposed method can produce robustness to this situation and obtain good detection results. As shown in Fig.8, there is a lot of garbage in the shade of trees but it can still be detected correctly with a high degree of confidence. In actual detection, the strong and weak degradations often occur, and our method can maintain a considerable accuracy.

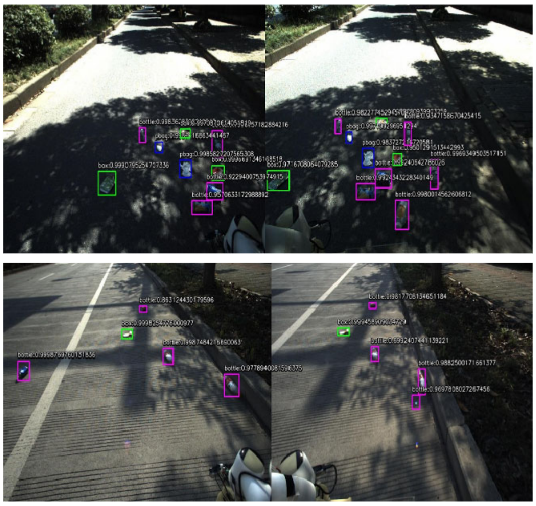


Fig.8 Actual detection results, where two images on the left and right represent the images collected by the left and right cameras

In this paper, we solve the problem of object detection in degraded images with unbalanced training samples. The dataset contains a large amount of daytime data and a small amount of nighttime data. In order to improve the detection accuracy of garbage on nighttime data while maintaining satisfactory detection accuracy of garbage on daytime data, we propose Mask Augmentation and Multi-Feature Attention Module. This method can also increase the robustness to degraded images. In most cases of degradation caused by strong light, shadows, and nighttime scenes, our method performs well. When compared with the benchmark model and other models in the experiment, our method significantly improves the detection accuracy of nighttime data and also performs well on daytime detection.

References

[1] Ren S, He K, Girshick R and J. Sun, IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 1137 (2015).

[2] Cai Z and Vasconcelos N, Cascade R-CNN: Delving into High Quality Object Detection, arXiv:1712.00726v1, 2018.

[3] Redmon J and Farhadi A, YOLOv3: An Incremental Improvement, arXiv:1804.02767, 2018.

[4] Tian Z, Shen C, Chen H and He T, Fcos: Fully Convolutional One-Stage Object Detection, IEEE International Conference on Computer Vision, 9627 (2019).

[5] Kong T, Sun F, Liu H, Jiang Y and Shi J, Foveabox: Beyond Anchor-Based Object Detector, arXiv:1904.03797, 2019.

[6] Yang Z, Liu S, Hu H, Wang L and Lin S, Reppoints: Point Set Representation for Object Detection, IEEE International Conference on Computer Vision, 9657 (2019).

[7] Proença P F and Simões P, TACO: Trash Annotations in Context for Litter Detection, arXiv:2003.06975, 2020.

[8] Lowe D G, International Journal of Computer Vision **60**, 91 (2004).

[9] Platt J, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998.

[10] Girshick R, Donahue J, Darrell T and Malik J, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, IEEE Conference on computer vision and pattern recognition, 580 (2014).

[11] Girshick R, Fast R-CNN, IEEE International Conference on Computer Vision, 1440 (2015).

[12] Redmon J, Divvala S, Girshick R and Farhadi A, You Only Look Once: Unified, Real-Time Object Detection, IEEE Conference on Computer Vision and Pattern Recognition, 779 (2016).

[13] Redmon J and Farhadi A, YOLO9000: Better, Faster, Stronger, IEEE Conference on Computer Vision and Pattern Recognition, 7263 (2017).

[14] Lin T Y, Goyal P, Girshick R, He K and Dollár P, Focal Loss for Dense Object Detection, IEEE International Conference on Computer Vision, 2980 (2017).

[15] Zhou X, Wang D and Krähenbühl P, Objects as Points, Computer Vision and Pattern Recognition, arXiv:1904.07850, 2019.

[16] Fan D P, Ji G P, Sun G, Cheng MM and L Shao, Camouflaged Object Detection, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[17] Fan Q, Zhuo W, Tang C K and Tai Y W, Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[18] Cao J, Cholakkal H, Anwer R M, Khan F S and Shao L, D2Det: Towards High Quality Object Detection and Instance Segmentation, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[19] Ioffe S and Szegedy C, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv:1502.03167, 2015.

[20] Maas A L, Hannun A Y and Ng A Y, Rectifier Nonlinearities Improve Neural Network Acoustic Models, ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 3 (2013).

[21] He K, Zhang X, Ren S and Sun J, Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition, 770 (2016).

[22] Neubeck A and Van Gool L, Efficient Non-Maximum Suppression, 18th International Conference on Pattern Recognition, 850 (2006).

[23] Lin M, Chen Q and Yan S, Network in Network, arXiv:1312.4400, 2013.

[24] Lin T Y, Maire M, Belongie S, Hays J and Zitnick C, Microsoft COCO: Common Objects in Context, European Conference on Computer Vision, 740 (2014).

[25] Deng J, Dong W, Socher R, Li L and Li F, Imagenet: A Large-Scale Hierarchical Image Database, IEEE Conference on Computer Vision and Pattern Recognition, 248 (2009).

[26] Bochkovskiy A, Wang C Y and Liao H Y M, YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv:2004.10934, 2020.