ISSN 2278-3091
Volume 12, No.3, May - June 2023
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse011232023.pdf
https://doi.org/10.30534/ijatcse/2023/011232023

A Scaling Factor Based Image Processing Strategy for Object Detection

Zhuohui Chen[1], Shunying Lin[2], ZeKai He[3], Ling Chen[*4]
[1]Macau University of Science and Technology, China, 2113950286@qq.com
[2]Hong Kong Baptist University, China, linshunying12@163.com
[3]Beijing Institute of Technology, Zhuhai, China, zekaihe@126.com
[4]Beijing Institute of Technology, Zhuhai, China, lingchensh@126.com

Received Date April 10, 2023 Accepted Date: May 15, 2023 Published Date: June 06, 2023

ABSTRACT
Classified management of domestic garbage is conducive to controlling pollution, protecting the environment, saving resources and achieving sustainable urban development. To automate domestic garbage classification and improve classification rate and processing capacity, this paper innovatively proposes an image processing strategy to detect domestic garbage objects using domestic garbage images as a dataset and YOLOv5 network. The network is then fine-tuned to achieve object detection of domestic garbage. Experimental results show that after using the image processing strategy, mAP@.5:.95 of the first-class (4-class) and second-class (104-class) networks on the basic test set is increased from 15.4% and 10.9% to 28.4% and 18.5%, respectively. This demonstrates the feasibility and effectiveness of the proposed image processing strategy. In addition, the image processing strategies presented in this paper have the potential to be applied in the domain of video recognition, including Sign Language Translation and Lip-reading Recognition.

Key words: Domestic Garbage Image Dataset, Image Processing Strategy, YOLOv5 Network

1. INTRODUCTION

Regulations on Municipal solid waste Management in Beijing was officially implemented on May 1st, 2020. The classified management of domestic waste is conducive to controlling pollution, protecting the environment, saving resources, promoting ecological civilization, and achieving sustainable urban development. Garbage classification is a reform of the traditional method of garbage collection and disposal, and a scientific management method for efficient garbage disposal. In the face of increasing garbage output and environmental deterioration, people can reduce garbage disposal to the maximum extent, realize the utilization of litter resources, and improve the quality of the living environment through the classification of waste management.

To help automatic classification of domestic garbage and upgrade classification rate and processing capacity, this paper proposes an image processing strategy using domestic garbage images as dataset and YOLOv5 network as domestic garbage detection network.

The organization of the paper is as follows. Section 2 reviews the two-stage object detection and single-stage object detection networks in detail. In Section 3, a domestic garbage image dataset is pre-processed as a benchmark dataset for domestic garbage detection networks. Section 4 proposes an image processing strategy including image combination method and image denoising algorithm, and describes the network structure and loss function of YOLOv5. Section 5 is devoted to the experimental implementation. The benchmark dataset and the YOLOv5 network are used to evaluate the proposed image processing strategy. Section 6 is a summary of the full paper. Overall, the contributions of this paper are as follows:

- Based on the rationality and standardization of data, this paper rationalizes and normalizes the image dataset of domestic garbage, and ensures the data quality of the dataset.
- This paper presents an image processing strategy, which enables the network to learn the image features with complex background, and improves the accuracy of household garbage detection.
- The experiment is carried out on the image dataset of domestic garbage, and it verifies the effectiveness of the method.

2. RELATED WORK

Lecun et al. [1] published a paper in IEEE in 1998, in which a convolutional-pooled-fully connected neural network structure was proposed for the first time. The seven-layer network was named LeNet-5, so Lecun also won the reputation of "the father of convolutional neural networks" [2]. The proposal of LeNet-5 enabled the successful commercialization of convolutional neural networks at that time, which were widely applied to the recognition tasks of postal codes, check numbers, etc. It can also be used to identify handwritten numbers and machine-printed character

pictures [3]. Convolutional neural networks are widely used in computer vision fields such as image recognition, object detection, and semantic segmentation.

There are two main branches of object detection, one is two-stage object detection method, including R-CNN [4], SPP-Net [5], Fast R-CNN [6], Faster R-CNN [7], R-FCN [8] and Mask R-CNN [9] while the other is single-stage object detection method, including SSD [10], RetinaNet [11], RefineDet [12] and YOLO [13]. YOLO is an object detection network proposed by Redmon in 2015 that performs only one forward propagation. It is the first single-stage object detection network. However, compared to the two-stage object detection system, YOLO generates more localization errors and lags in accuracy, especially for small target object detection. At the same time, YOLO imposes spatial constraints on bounding box predictions, since each grid cell predicts only two bounding boxes and can only have one class, which limits the number of neighboring targets that YOLO can predict. Therefore, the effect of using YOLO to detect birds, crowds and motorcades was not ideal [14]. In the same year, Liu et al. [10] gave the SSD method. SSD has absorbed the idea of YOLO fast detection, combined the advantages of Faster R-CNN and Region Proposal Network (RPN) and improved the processing method of multi-size targets, instead of only using top-level feature maps for prediction. Experimental results show that SSD performs slightly better than Faster R-CNN on the Pascal VOC 2007 dataset, achieving 6.6 times faster than Faster R-CNN, thus achieving a comprehensive performance improvement over two-stage object detection.

At the end of 2016, Redmon et al. [15] introduced the YOLOv2 network. Under the processing speed of 67 frames per second, YOLOv2 obtained 76.8% mAP on Pascal VOC 2007 dataset. At a processing speed of 40 frames per second, YOLOv2 achieved 78.6% mAP, which was better than the SSD method.

In 2020, Ultralytics et al. released the source code of the YOLOv5 network. The network structure of YOLOv5 is the same as that of YOLOv4 [16]. They all use Crossstage Partial Networks (CSP Net) as the backbone and use Path Aggregation Network (PANet) for information transmission. The CSPNet was proposed by Wang et al. [17] to solve the problem that the network structure needs a lot of reasoning calculation. They attributed the problem to repeated gradient information in network optimization. The experimental results show that the AP_{50} on MS COCO object detection dataset using CSPNet significantly outperforms other methods. PANet is a network to promote information circulation proposed by Liu et al. [18]. It won first prize in the COCO2017 challenge instance segmentation task without mass training and second prize in the object detection task. Unlike YOLOv4, which uses the Mish function as the convolutional layer activation function, YOLOv5 uses the SiLU function as the convolutional layer activation function. SiLU function is an activation function of neural network function approximation proposed by Elfwing et al. [19]. Experimental results show that SiLU activation function significantly outperforms ReLU activation function, and a Sarsa algorithm with SiLU activation function outperforms DQN network. Regarding the loss function, YOLOv5 still uses the CIoU function of YOLOv4 as the loss function for bounding box regression. The CIoU function is a generalized version of IoU introduced by Zheng et al. [20], which is used to better distinguish between easy and difficult regression bounding boxes. Experimental results show that the performance of object detection networks with CIoU bounding box regression loss function can be significantly improved without sacrificing inference efficiency.

While these methods achieve decent results for object detection, their performance is inferior for multi-object detection tasks with complex backgrounds. Therefore, this paper puts forward an image processing strategy, which combines the images from the training set and denoises the images from the test set to enable the network to learn the image features with complex backgrounds. The YOLOv5 network is used to detect objects in the domestic garbage image dataset. Based on this, image processing strategies have been added to improve the accuracy of domestic waste detection.

3. DOMESTIC GARBAGE IAMGE DATASET

This section first introduces the image dataset of domestic garbage and then processes the images of domestic garbage from the perspective of rationality and normalization to ensure the data quality of the dataset.

3.1 Dataset Introduction

The domestic garbage image dataset used in this paper was provided by 2021 Zhuhai Open Data Innovation Apps Contest. The dataset contains 6639 images and 6613 label information including 6629 images in the training set and 10 images in the testing set. The images were divided into 4 first-class categories and 104 second-class categories. The first-class categories consisted of other waste, harmful waste, recyclables, and kitchen waste. The second-class categories are respectively Disposable fast food boxes, Fruit pulp, Fruit peel, Tea residue, Cabbage root, Eggshell, Fishbone, Power bank, Pack, Cosmetics bottles, Plastic toys, Plastic bowls and basins, Plastic hangers, Fouling plastic, Express paper bags, Plug wire, Old clothes, Cans, Pillow, Plush toys, Shampoo bottle, Glass, Leather shoes, Chopping block, Cigarette butts, Cardboard box, Spice bottle, Wine bottles, Metal food cans, Wok, Cooking oil drum, Beverage bottles, Dry batteries, Ointment, Expired medications, Toothpick, Break pots and dishes, Chopsticks, Leftovers, Big bones, LED flashlight, Health supplement bottle, Eight treasure porridge, Ice cream, Rock sugar gourd, Stool, Shoulder bag, Masks, Apron, Globe, Sweet potato, Nut, Biscuits, Milk tea cups, Desiccant, Cell phone, Nail polish, Warm the baby, Insecticide, Jelly, Walnut, Orange, Leftovers, Intestines, Correction, Thermometer, Wet

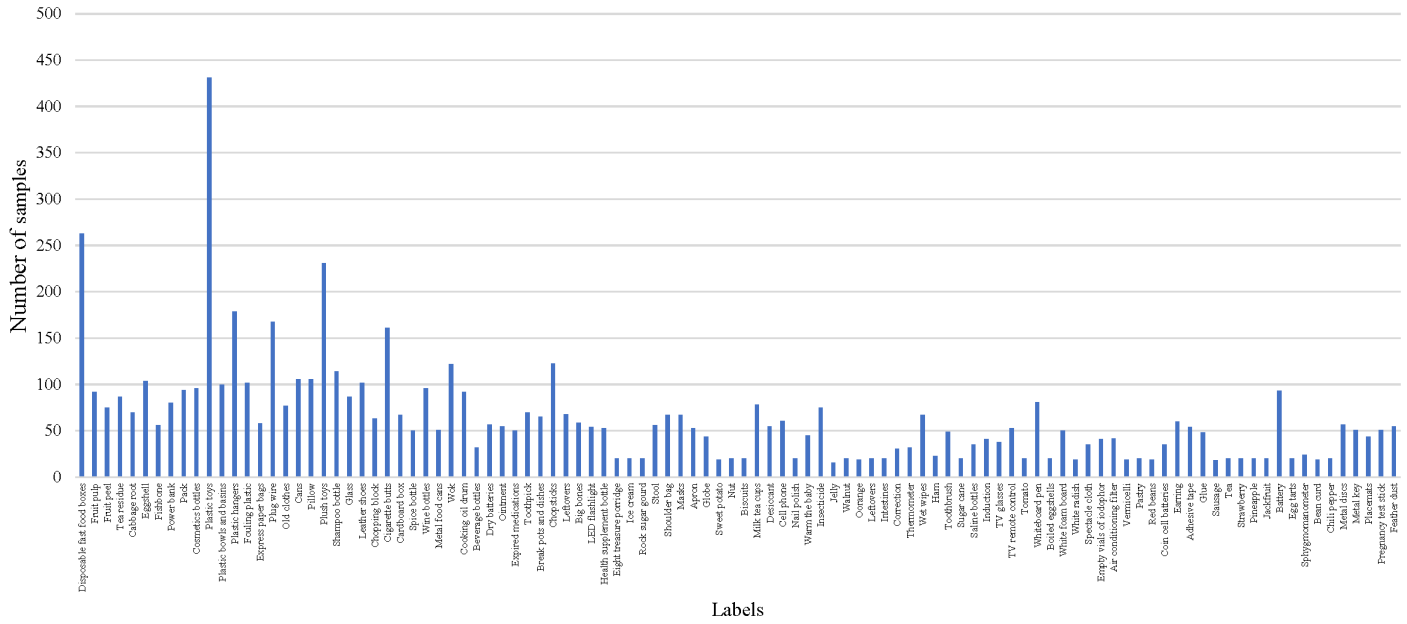


Figure 1: Bar Chart of Second-class Category Quantity of Domestic Garbage Dataset

wipes, Ham, Toothbrush, Sugar cane, Saline bottles, Induction, TV glasses, TV remote control, Tomato, Whiteboard pen, Boiled eggshells, White foam board, White radish, Spectacle cloth, Empty vials of iodophor, Air conditioning filter, Vermicelli, Pastry, Red beans, Coin cell batteries, Earring, Adhesive tape, Glue, Sausage, Tea, Strawberry, Pineapple, Jackfruit, Battery, Egg tarts, Sphygmomanometer, Bean curd, Chili pepper, Metal discs, Metal key, Placemats, Pregnancy test stick and Feather dust. The image numbers for other waste, kitchen waste, 3338, and 618, respectively. The image numbers in the second category are shown in Figure 1. The location information includes 4 location information of the items, namely the top-left and bottom-right coordinates of the items, namely $x_{\min}$, $y_{\min}$, $x_{\max}$ and $y_{\max}$.

3.2 Data Preprocessing

A. Data Rationalization

After statistics, it is found that there are the following five problems in the image and label information among the datasets: (1) There is a sample whose first-class category "harmful waste" is erroneously written as "hazardous substance"; (2) The location information of some samples is less than 0; (3) The location information of some samples is beyond the boundary; (4) The location information of some samples is missing; (5) There are two images that cannot be read.

In this regard, the following actions are taken to rationalize the data: (1) Modify the first-class classification of "hazardous substances" to "harmful waste"; (2) Take the location information which is less than 0 as 0; (3) Directly taking the location information beyond the boundary as the boundary value; (4) Delete 28 samples which missing position information; (5) Delete 2 samples which failed in image reading. After rationalization, the sample size of the dataset is 6593, that is, it contains 6593 images and 6583 label information.

B. Data Normalization

In order to make the domestic garbage dataset capable of being trained by the YOLOv5 network, it is necessary to normalize the label information of the dataset, and the transformation formulas are as follows:

$$x_{\text{center}} = (x_{\max} - 0.5w_{\text{box}}) / w \quad (1)$$

$$y_{\text{center}} = (y_{\max} - 0.5h_{\text{box}}) / h \quad (2)$$

$$w_{\text{box}} = x_{\max} - x_{\min} \quad (3)$$

$$h_{\text{box}} = y_{\max} - y_{\min} \quad (4)$$

$$w'_{\text{box}} = w_{\text{box}} / w \quad (5)$$

$$h'_{\text{box}} = h_{\text{box}} / h \quad (6)$$

where w and h are the width and height of the image.

4. METHODOLOGY

This section first introduces an image processing strategy that can improve the accuracy of object detection for domestic garbage. Then, the YOLOv5 network is used to build first-class and second-class classification networks for domestic waste to achieve object detection for domestic waste.

4.1 Image Processing Strategy

A. Image Combination Method

First, 20% of the training set is selected as the validation set and the rest as the new training set. That is, the training set has a sample size of 5266, the validation set has a sample size of

1317, and the test set has a sample size of 10. Then, by looking at the test set images of domestic garbage, it is found that most of the images contain two or three items. Therefore, according to the characteristics of the test set images, the image combination method is used, that is, two or three images of the training set and the validation set are randomly and repeatedly combined into a large image, and a total of 2108 new training set images and 527 validation set images are synthesized, as shown in Figure 2. The combined image sizes are all 489n×458, where 489 is the average of the widths of all the images in the training set and the validation set, 458 is the high average of all the images in the training set and the validation set, and n is the number of randomly combined images.

Since the images are combined, it is necessary to re-label the combined images. The transformation formulas for the location information of objects in the combined image are as follows:

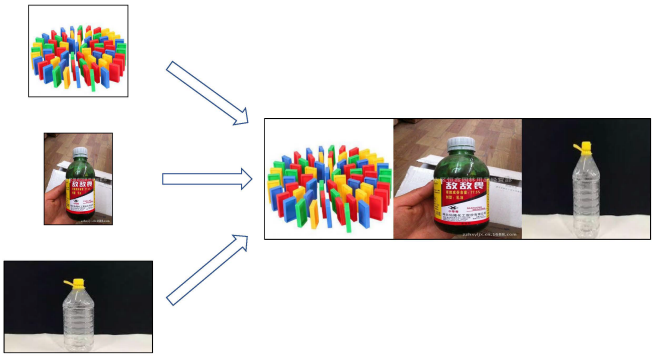


Figure 2: Image Combination Method

$$r_w = \frac{w}{w'} \tag{7}$$

$$r_h = \frac{h}{h'} \tag{8}$$

$$x'_{\text{center}} = \frac{x_{\text{center}} w\overline{w}}{r_w (w')^2} + (i-1)\frac{\overline{w}}{w'} \tag{9}$$

$$y'_{\text{center}} = \frac{y_{\text{center}} h}{r_h h'} \tag{10}$$

$$w''_{\text{box}} = \frac{w'_{\text{box}}\, w\overline{w}}{r_w (w')^2} \tag{11}$$

$$h''_{\text{box}} = \frac{h'_{\text{box}}\, h}{r_h h'} \tag{12}$$

where w and h are the width and height of the pre-combined image, w' and h' are the width and height of the post-combined image, x'_{center}, y'_{center}, w''_{box} and h''_{box} are the four location information after the transformation, and i is the i-subgraph in the combined image. Simplify formulas (7) to (12), and find that after transformation, y'_{center} and h''_{box} are unchanged, as follows:

$$x'_{\text{center}} = \frac{x_{\text{center}} w\overline{w}}{r_w (w')^2} + (i-1)\frac{\overline{w}}{w'} \tag{13}$$

$$y'_{\text{center}} = y_{\text{center}} \tag{14}$$

$$w''_{\text{box}} = \frac{w'_{\text{box}}}{n} \tag{15}$$

$$h''_{\text{box}} = h'_{\text{box}} \tag{16}$$

where n is the number of randomly combined images.

B. Image Denoising Algorithm

After observing the image of the training set of domestic garbage, it is noticed that there is a “wallet” in the test set, but there is no “wallet” in the second-class classification of domestic garbage, so it should be removed as shown in Figure 3.

(a) Before Elimination (b) After Elimination

Figure 3: Eliminate the Nonexistent Category Images

Moreover, most of the image backgrounds in the test set are monochromatic, but the images in the test set have environmental disturbances such as light sources, desk gaps and other factors. Therefore, OpenCV techniques were used to remove background colors from the test set images. The images before and after processing are shown in Figure 4.

(a) Before Removal (b) After Removal

Figure 4: Eliminate the Image Background Color

Next, the object outline of the test set image is extracted by the OpenCV technique, and the extracted object set is denoised. The pseudocode of the scaling factor based image denoising algorithm is given in Algorithm 1.

Algorithm 1 Scaling Factor Based Image Denoising Algorithm

Require:

The extracted object image set X, the extracted object image height h, the extracted object image width w, the average value $\overline{h}$ of the heights of the training set and the validation set images, and the average value $\overline{w}$ of the widths of the training set and the validation set images

Ensure:

Normalized collection of object images X'

1: **for** $x \in X$ **do**
2: **if** the height of the object image is greater than the width **then**
3: compress the height of the object image to $\overline{h}$ and the width to $w\overline{h}/h$
4: **if** the width of the object image is larger than $\overline{w}$ **then**
5: compress the width of the object image to $\overline{w}$
6: **else if** the width of the object image is shorter than $\overline{w}$ **then**
7: expand all 0-column vectors on the left and right sides of the object image until the image width is $\overline{w}$
8: **end if**
9: **else if** the height of the object image is equal to the width **then**
10: change the height of the object image to $\overline{w}$, the width to $\overline{w}$, and expand all the 0-line vectors on the top and bottom sides of the object image until the image height is $\overline{h}$
11: **else**
12: compress the width of the object image to $\overline{w}$, and change the width to $h\overline{w}/w$
13: **if** the height of the object image is greater than $\overline{h}$ **then**
14: compress the height of the object image to $\overline{h}$
15: **else if** the height of the object image is shorter than $\overline{h}$ **then**
16: expand all the 0-line vectors on the top and bottom sides of the object image until the height of the object image is $\overline{h}$
17: **end if**
18: **end if**
19: **end for**

Subsequently, 25 random monochromes were set to change the background color of the 25 items in the test set mages. Next, it is concatenated horizontally, and the result is shown in Figure 5. Finally, the Binary object detection and labeling software was used to annotate the test set images with the label information of the YOLOv5 network, resulting in a total of 10 labels.

(a) Before Change (b) After Change

Figure 5: Comparison Before and After Image Background Change

4.2 Network Structure

The YOLO network is the earliest single-stage object detection method, and it is the first method to achieve real-time object detection, which can reach a speed of 45 frames. Moreover, the mAP of the YOLO network is twice or even higher than that of other real-time detection systems. The YOLO network provides new insight into the detection speed of deep learning based methods for object detection. YOLOv5, a newly released network in the YOLO series, is slightly weaker than YOLOv4 in performance, but much stronger than YOLOv4 in terms of flexibility and speed, and has great advantages in the rapid deployment of models. In this paper, the YOLOv5 network is used to segment and identify images of domestic garbage. First, the network structure of YOLOv5 is shown in Figure 6.

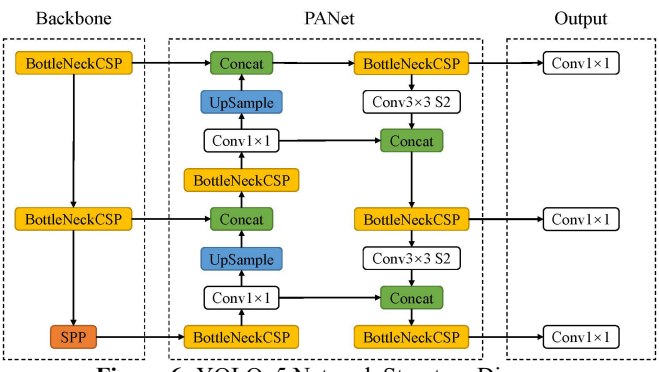


Figure 6: YOLOv5 Network Structure Diagram

Among them, BottleNeckCSP is the bottleneck residual module, which combines BottleNect and CSP and is the main structure for residual learning. BottleNect is the bottleneck layer, which is used to reduce the number of channels. SPP is the SPP-Net, which is used to keep the shape of the pooled features unchanged. Conv is the convolutional layer used to extract the features of the image, where S2 denotes the convolution step size is 2. Concat is the splicing layer. UpSample is the upsampling layer, which is used to enlarge the image. PANet is a path aggregation network.

4.3 Loss Function

The YOLOv5 network uses the CIoU function to compute the bounding box regression loss, the log-binomial cross entropy function to compute the objective loss, and the log-binomial cross entropy function to compute the object classification loss, so the total loss is formulated as follows:

$$L = bs(L_{\text{box}} + L_{\text{obj}} + L_{\text{cls}}) \tag{17}$$

where *bs* stands for batch training size, L_{box} for bounding box regression loss, L_{obj} for objectivity loss, and L_{cls} for object classification loss. Box regression calculates the loss through the CIoU function, which is formulated as follows:

$$L_{\text{box}} = 1 - \text{IoU} + \frac{\rho(\mathbf{p}, \hat{\mathbf{p}})}{c^2} + \alpha V \tag{18}$$

$$\alpha = \begin{cases} 0, & \text{if IoU} < 0.5, \\ \dfrac{V}{1 - \text{IoU} + V}, & \text{if IoU} \geq 0.5. \end{cases} \tag{19}$$

$$V = \frac{4}{\pi^2}(\arctan\frac{w''_{\text{box}}}{h''_{\text{box}}} - \arctan\frac{\hat{w}''_{\text{box}}}{\hat{h}''_{\text{box}}})^2 \tag{20}$$

IoU denotes Intersection over Union, ρ denotes the Euclidean distance function, $\mathbf{p}=[x'_{\text{center}}, y'_{\text{center}}]^T$, $\hat{\mathbf{p}}=[\hat{x}'_{\text{center}}, \hat{y}'_{\text{center}}]^T$ denotes the central point of the bounding box, and *c* denotes the diagonal length of the bounding box.

The formula of the objective loss function is as follows:

$$L_{\text{obj}}(\text{IoU}, \hat{\text{IoU}}) = -\frac{1}{N}\sum_{i=1}^{N}[\text{IoU}_i \log(\sigma(\hat{\text{IoU}}_i)) + (1 - \text{IoU}_i)\log(1 - \sigma(\hat{\text{IoU}}_i))] \tag{21}$$

The formula of the object classification loss function is as follows:

$$L_{\text{cls}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(\hat{y}_i)) + (1 - y_i)\log(1 - \sigma(\hat{y}_i))] \tag{22}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{23}$$

where *N* is the sample size and *y* is the label.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the image processing strategy is first used in experiments, and the network parameters are fine-tuned to improve the performance of the network. Then, the performance of the network is evaluated by the evaluation metrics, confusion matrix, and classification reports of the network on the test set.

5.1 Experimental Parameters and Environment Settings

For the network training process in this paper, the hardware environment uses an Intel Core i5-10200H CPU processor, 2.40GHz 16GB memory, and a single NVIDIA GeForce RTX 2060 graphics processing unit. The Windows 10 software system was used as the operating system and the network was trained and tested on the framework of PyTorch 1.80. To improve the generalization ability of the model, the parameters in Table 1 are used and a pre-trained small-scale YOLOv5 network is chosen to train on the domestic garbage image dataset.

Table 1: Network Training Parameters

Epochs	Batch Size	Optimizer	Validation Set Size
20	16	SGD	0.1

Among them, the batch size is 16 and the optimization algorithm SGD with stochastic gradient descent is chosen as the optimizer of the network. The detailed parameters are listed in Table 2.

Table 2: SGD Optimizer Parameters

Learning Rate	Momentum	Weight decay	Nesterov Momentum
0.01	0.937	5×10^{-4}	enable

The momentum formula is as follows:

$$v_{t+1} = \mu v_t - \alpha \nabla L(w_t) \tag{24}$$

$$w_{t+1} = w_t + v_{t+1} \tag{25}$$

where μ ($\mu \in [0,1]$) is the momentum factor, α ($\alpha > 0$) is the learning rate and *w* is the weight. If the Nesterov momentum is enabled, the momentum formula is updated to the NAG formula:

$$v_{t+1} = \mu v_t - \alpha \nabla L(w_t + \mu v_t) \tag{26}$$

$$w_{t+1} = w_t + v_{t+1} \tag{27}$$

In addition, L_2 regularization reduces the network weights by adding a penalty term to the loss function of the network. The updated formula for the loss function is given as follows:

$$L(w) = \frac{\lambda}{2N} \| w \|^2 \tag{28}$$

where λ is the attenuation coefficient. The YOLOv5 network also sets the attenuation rate of the learning rate, which is updated by the following formula:

$$\alpha = 1 - \frac{\alpha(1-\alpha_f)}{E_p - 1} + \alpha_f \tag{29}$$

where E_p refers to the number of training epochs of the network, and α_f is the value at which the ensemble learning rate stops updating.

5.2 Evaluation Metrics

In this paper, four evaluation indicators, precision rate, recall rate, mAP@.5, and mAP@.5:.95 are selected as the evaluation indexes of the model. The formulas of precision rate *P* and recall rate *R* are as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{30}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{31}$$

where TP stands for True Positive, which is the correct number of positive samples identified; FP stands for False Positive and is the number of positive samples that have been incorrectly identified; FN stands for False Negative, and denotes the number of false negative samples that have been incorrectly identified.

The calculation formula of mean Average Precision (mAP) is as follows:

$$\mathrm{mAP} = \frac{\sum_{c=1}^{K} \mathrm{AP}_c}{K} \tag{32}$$

where *K* represents the total number of categories, and AP represents the area under the precision and recall curve (PR curve) under category *c*. mAP@.5 is the calculated value of IoU under the condition of 0.5; mAP@.5:.95 is the value obtained by taking the IoU from 0.5 to 0.95, one mAP is calculated every 0.05, and the ten mAPs are finally average.

5.3 Model Training

Firstly, the original training set and validation set images are used to train the network, and it is concluded that the training time of the first-class and second-class classification networks is 8178 seconds and 7648 seconds respectively, and the mAP@.5:.95 of the networks on the validation set has individual scores of 0.219 and 0.123.

Then, the combined training set and validation set images are used to train the network, and it is concluded that the training time of the first-class and second-class classification networks is 3852 seconds and 3603 seconds, respectively, and the mAP@.5:.95 of the networks on the validation set has corresponding values of 0.362 and 0.146. It can be found that using the image combination method not only improves the network performance but also reduces the time-consuming network training. The original test set is then loaded into the network for testing, and the evaluation metrics of the first-class classification network on the test set are shown in Table 3.

Table 3: Evaluation Metrics of the First-class Classification Network on the Test Set

Domestic Garbage Category	*P*	*R*	mAP @.5	mAP@ .5:.95
All Categories	0.488	0.433	0.400	0.233
Harmful Waste	0.505	0.750	0.681	0.454
Kitchen Waste	1.000	0.385	0.575	0.245
Recyclables	0.460	0.500	0.303	0.213
Other Waste	0.200	0.125	0.043	0.019

Each evaluation metric of the second-class classification network on the test set is shown in Table 4.

Table 4: Evaluation Metrics of the Second-class Classification Network on Test Set

Domestic Garbage Category	*P*	*R*	mAP @.5	mAP@ .5:.95
All Categories	0.735	0.253	0.264	0.176

Finally, the denoised test set is put into the network for testing, and the evaluation metrics of the first-level classification network on the normalized test set are given in Table 5.

Table 5: Evaluation Metrics of the Second-class Classification Network on the Normalized Test Set

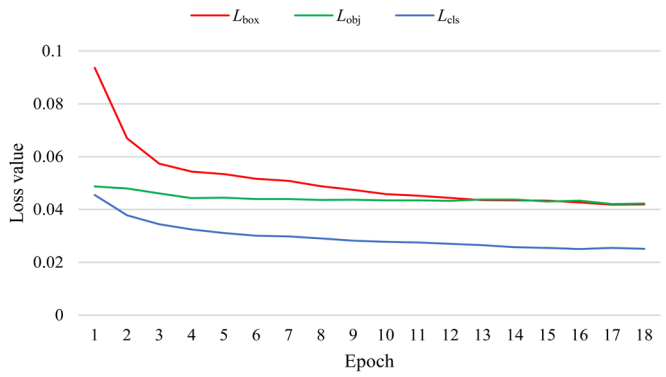
Domestic Garbage Category	*P*	*R*	mAP @.5	mAP@ .5:.95
All Categories	0.545	0.483	0.474	0.284
Harmful Waste	0.635	0.750	0.758	0.555
Kitchen Waste	0.638	0.125	0.309	0.111
Recyclables	0.320	0.700	0.489	0.336
Other Waste	0.587	0.358	0.340	0.135

Each evaluation metric of the second-class classification network on the denoised test set is shown in Table 6.

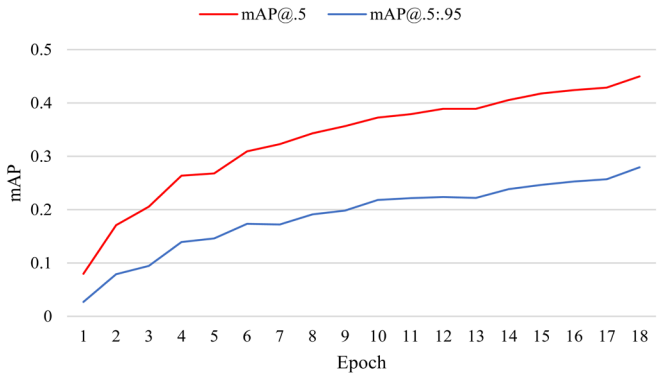
Table 6: Evaluation Metrics of the Second-class Classification Network on Normalized Test Set

Domestic Garbage Category	*P*	*R*	mAP @.5	mAP@ .5:.95
All Categories	0.824	0.192	0.281	0.185

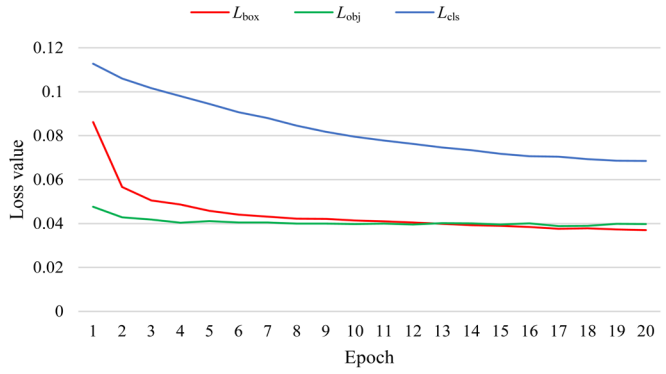
Obviously, after denoising the test set images, all the evaluation indicators of the network are improved. Finally,



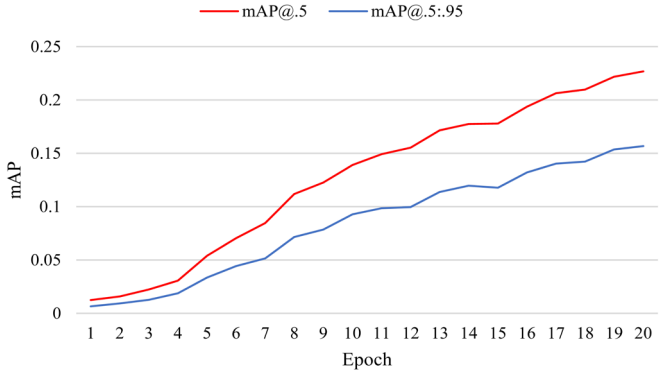
(a) Loss Value of the First-class Classification Network



(b) mAP of the First-class Classification Network



(c) Loss Value of the Second-class Classification Network



(d) mAP of the Second-class Classification Network

Figure 7: Variation Curves of Each Indicators of the Network During Training

fine-tune the parameters of the network. The fine-tuned parameters of the first-class and second-class classification networks are shown in Table 7.

Table 7: Evaluation Metrics of the Second-class Classification Network on Normalized Test Set

Model	Epochs	Batch Size	Weight decay
First-class Classification	**18**	16	5×10^{-4}
Second-class Classification	20	**8**	**5×10^{-6}**

After fine-tuning the parameters of the network, it is found that the mAP@.5:.95 of the first-class classification network on the validation and test set has a score of 0.279 and 0.404 respectively, while the second-class classification network has scores of 0.157 and 0.275, respectively. Also, Figure 7 shows the loss values and mAP curves for the first-class and second-class classification networks during the training process. It can be concluded that the boundary box regression loss of the first-class and second-class classification networks is not much different from the objective loss, while the loss of the first-class classification network is much lower than that of the second-class classification network, because the number of first-class categories is less than that of second-class categories. In addition, the mAP of the first-class and second-class classification networks all rose rapidly at the beginning, then stabilized, the training epochs were enough, and the network had reached a stable state.

5.4 Model Evaluation

A. Ablation Study

First, an ablation study using the method presented in this study was performed and the results are shown in Table 8. It can be seen that using the training set image combination method improves the mAP@.5:.95 of the first-class classification network by 7.9% and that of the second-class classification network by 6.7%, which is quite remarkable. Using the image denoising algorithm also can further improve the performance of the first-class and second-class classification networks. Therefore, it proves that the proposed image processing strategy is very effective and can significantly improve the accuracy of object detection on domestic garbage images.

Second, the confusion matrix of the first-class classification network is plotted on the test set and evaluated, as shown in Figure 8. The ideal output result of the confusion matrix is that the value of the same category is 1 and the value of different categories is 0. The results show that the confusion matrix value of the first-class classification network is 0.6 for

Table 8: Results of Ablation Study

Image Processing Strategy		Fine-tune	MAP@.5:.95 of the First-class Classification Network	MAP@.5:.95 of the Second-class Classification Network
Image Combination Method	Image Denoising Algorithm			
			0.154	0.109
✓			0.233	0.176
✓	✓		0.284	0.185
✓	✓	✓	0.404	0.275

recyclables, 0.25 for harmful waste and kitchen waste, and less than 0.25 for other waste. It can be seen that the proposed model has the best object detection performance on recyclable images, average performance on harmful waste and kitchen waste images, and weak performance on other waste images.

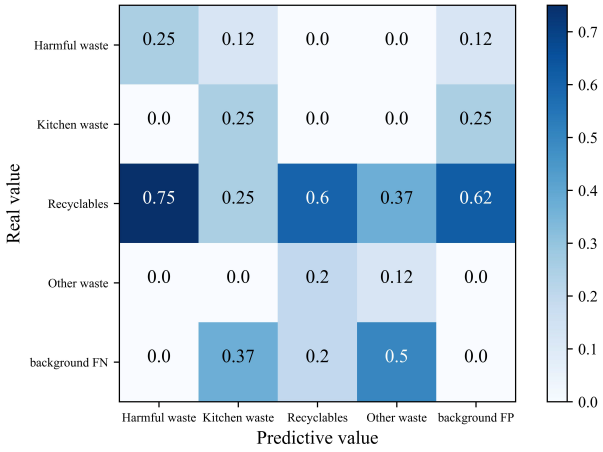


Figure 8: Confusion Matrix

Third, the evaluation indicator results of the network are output by calculating *P*, *R*, mAP@.5, and mAP@.5:.95. In this way, the network is evaluated and the results of the evaluation indicators for the first-class classification network in all categories are shown in Table 9. The results show that the network performs well for object detection on images of hazardous waste and recyclables, but mediocre on images of kitchen waste and other waste. Moreover, the average mAP@.5:.95 of the first-class classification network reaches 0.404, and the overall object detection effect of the network is good.

Table 9: Evaluation Results of First-class Classification Network

Domestic Garbage Category	*P*	*R*	mAP @.5	mAP@ .5:.95
All Categories	0.545	0.483	0.474	0.284
Harmful Waste	0.635	0.750	0.758	0.555
Kitchen Waste	0.638	0.125	0.309	0.111
Recyclables	0.320	0.700	0.489	0.336
Other Waste	0.587	0.358	0.340	0.135

Since there are too many second-class categories, the confusion matrix and classification report of the second-class classification network are not shown here. The accuracy rate, recall rate, mAP@.5, and mAP@.5:.95 of the two-class classification network in all categories are 0.488, 0.479, 0.391, and 0.275, respectively, and the overall object detection effect of the network is average.

B. Error Analysis

According to statistics, the details of the domestic garbage images of the first-class and second-class categories are shown in Table 10. Besides, the images of other waste, kitchen waste, recyclables, and harmful waste in the test set are 4, 8, 10, and 8, respectively. Combining Figure 1 and Table 10, it can be seen that there is a sample imbalance in the domestic garbage dataset, which leads to the low performance of the first-class and second-class classification networks. Furthermore, the imbalance in the samples of the second-class images is more severe than that of the first-class images. The average number of training validation sets for the first-class images is about 1657, while the average number of training validation sets for the second-class images is only about 60. Additionally, the number of test sets of the second-class images is as high as 26, which significantly reduces the performance of the second-class classification network.

Table 10: Details of Domestic Garbage Images of First-class and Second-class

Class	Samples		Category		mAP@.5:.95
	Others	Test	Others	Test	
First	6629	10	4	4	0.404
Second	6629	10	104	26	0.275

6. CONCLUSION

Object detection from domestic garbage has long been an urgent problem to be addressed. First, the image of domestic garbage is used as the benchmark dataset and the data is processed from the perspective of rationality and normalization. Second, an innovative image processing strategy is proposed, which consists of a training set image combination strategy and a test set image denoising algorithm. Next, the YOLOv5 network is used to build a first-class classification network and a second-class classification

network, respectively, to detect objects in domestic garbage images. The image processing strategy is then used in the experiments, and it is shown that the accuracy of the first-class classification network and the second-class classification network is 28.4% and 18.5% on the test set, respectively. Experimental results show that the proposed image processing strategy unifies the classification criteria of domestic garbage images and obtains the best object detection performance of the network. Finally, the network is fine-tuned and evaluated, and the mAP@.5:.95 of the first-class and second-class networks on the test set is 40.4% and 27.5%, respectively. The object detection of domestic garbage is implemented, and it is concluded that the overall performance of the first-class and second-class classification models is better for domestic garbage object detection. The proposed image processing strategy is mature. Additionally, this can be applied in the domain of video recognition, including sign language translation and recognition, as well as lip-reading translation.

In addition, the imbalance of samples in the domestic garbage datasets leads to the low quality of the first-class and second-class classification networks. After the error analysis of the first-class and second-class classification networks, it is found that in the future, it is necessary to crawl the category images which are seriously missing, so as to eliminate the influence of sample imbalance on the network, and then carry out object detection on domestic garbage. At the same time, further research on the object detection network is needed to further optimize the network so that it can better adapt to domestic garbage in different areas and environments and achieve automatic classification and delivery of domestic garbage. Moreover, it is also necessary to deploy the network to mobile phones and Internet platforms to promote the development of intelligent applications for domestic garbage classification and innovative grassroots social governance.

ACKNOWLEDGMENT

This work is supported by research grants from the Guangdong Science and Technology Innovation Strategy Special Fund "Research on Sign Language Recognition System for Deaf and Mute People Based on Deep Learning" (Grant No. pdjh2022a0706).

REFERENCES

1. Y. Lecun, and L. Bottou. **Gradient-based learning applied to document recognition**, in *Proc. IEEE*, 1998, pp. 2278-2324.
2. W. Lu. ***Deep Learning Notes***; Beijing: Peking University Press, 2020, pp. 21, 39.
3. L. Long. ***TensorFlow Deep Learning – In-depth understanding of AI algorithm design***; Beijing: Tsinghua University Press, 2020, pp. 132, 229.
4. R. Girshick, J. Donahue, and T. Darrell. **Rich feature hierarchies for accurate object detection and semantic segmentation**, in *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580-587.
5. K. He, X. Zhang, S. Ren, and J. Sun. **Spatial pyramid pooling in deep convolutional networks for visual Recognition**, in *Proc. 2014 European Conference on Computer Vision (ECCV)*, 2014, pp. 346-361.
6. R. Girshick. **Fast R-CNN**, in *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440-1448.
7. S. Ren, K. He, R. Girshick, and J. Sun. **Faster R-CNN: towards real-time object detection with region proposal networks**, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June 2017.
8. J. Dai, Y. Li, K. He, and J. Sun. **R-FCN: object detection via region-based fully convolutional networks**, in *Proc. 2016 International Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 379-387.
9. K. He, G. Gkioxari, P. Dollár, and R. Girshick. **Mask R-CNN**, in *Proc. 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980-2988.
10. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. **SSD: single shot multiBox detector**, in *Proc. 2016 European Conference on Computer Vision (ECCV)*, 2016, pp. 21-37.
11. T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. **Focal loss for dense object detection**, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 2, pp. 318-327, Feb 2020.
12. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. **Single-shot refinement neural network for object detection**, in *Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
13. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. **You Only Look Once: unified, real-time object detection**, in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788.
14. P. Du, M. Chen, T. Su. ***Deep Learning and Object Detection***; Beijing: Publishing House of Electronics Industry, 2020, pp. 153, 155.
15. J. Redmon, and A. Farhadi. **YOLO9000: better, faster, stronger**, in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
16. A. Bochkovskiy, C. Wang, and H. Liao. **YOLOv4: optimal speed and accuracy of object detection**, in *Proc. 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
17. C. Wang, H. M. Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh. **CSPNet: a new backbone that can enhance learning capability of CNN**, in *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1571-1580.
18. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. **Path aggregation network for instance segmentation**, in *Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8759-8768.
19. S. Elfwing, E. Uchibe, and K. Doya. **Sigmoid-weighted linear units for neural network function**

approximation in reinforcement learning, *Neural Networks*, Vol. 107, pp. 3-11, Nov 2018.

20. Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo. **Enhancing geometric factors in model learning and inference for object detection and instance segmentation**, *IEEE Transactions on Cybernetics*, Vol. 52, No. 8, pp. 8574-8586, Aug 2022.