Hindawi
Complexity
Volume 2021, Article ID 1055604, 14 pages
https://doi.org/10.1155/2021/1055604

Research Article

A Garbage Detection and Classification Method Based on Visual Scene Understanding in the Home Environment

Yuezhong Wu,[1,2] **Xuehao Shen**,[1] **Qiang Liu**,[2,3] **Falong Xiao**,[1] **and Changyun Li**[2,3]

[1]*College of Railway Transportation, Hunan University of Technology, Zhuzhou 412007, China*
[2]*College of Computer Science, Hunan University of Technology, Zhuzhou 412007, China*
[3]*Intelligent Information Perception and Processing Technology Hunan Province Key Laboratory, Zhuzhou, China*

Correspondence should be addressed to Qiang Liu; liuqiang@hut.edu.cn and Changyun Li; lichangyun@hut.edu.cn

Received 20 August 2021; Revised 14 September 2021; Accepted 2 November 2021; Published 26 November 2021

Academic Editor: Xuyun Zhang

Garbage classification is a social issue related to people's livelihood and sustainable development, so letting service robots autonomously perform intelligent garbage classification has important research significance. Aiming at the problems of complex systems with data source and cloud service center data transmission delay and untimely response, at the same time, in order to realize the perception, storage, and analysis of massive multisource heterogeneous data, a garbage detection and classification method based on visual scene understanding is proposed. This method uses knowledge graphs to store and model items in the scene in the form of images, videos, texts, and other multimodal forms. The ESA attention mechanism is added to the backbone network part of the YOLOv5 network, aiming to improve the feature extraction ability of the network, combining with the built multimodal knowledge graph to form the YOLOv5-Attention-KG model, and deploying it to the service robot to perform real-time perception on the items in the scene. Finally, collaborative training is carried out on the cloud server side and deployed to the edge device side to reason and analyze the data in real time. The test results show that, compared with the original YOLOv5 model, the detection and classification accuracy of the proposed model is higher, and the real-time performance can also meet the actual use requirements. The model proposed in this paper can realize the intelligent decision-making of garbage classification for big data in the scene in a complex system and has certain conditions for promotion and landing.

1. Introduction

In recent years, as the global garbage production has shown a cliff-like growth, my country has also introduced a series of policies. The latest revision of the "Law of the People's Republic of China on the Prevention and Control of Environmental Pollution by Fixed Wastes" in 2020 requires that local people's governments at or above the county level should speed up the establishment of a domestic waste management system for classified release, recycling, transportation, and treatment. At this stage, the garbage classification is mainly concentrated in fixed places in the outdoor public environment. There are problems such as high labor intensity, low sorting efficiency, and poor working environment. In fact, the garbage classification in the home environment can really solve the problem from the source. However, because the people's awareness of classification is not strong, the classification is troublesome, and there are many types of garbage; people seldom actually throw garbage in categories. In recent years, home service robots have attracted widespread attention. Among them, sweeping robots are the first products to realize industrialization and have entered the consumer market widely. Although the sweeping robots currently on the market have basic functions such as path planning [1, 2], automatic charging, and automatic obstacle avoidance, their intelligence is still not high. Although a simple path planning function is added to the cleaning process, the cleaning process is blind. No matter whether there is garbage in the working path that needs to be processed, the cleaning action will be performed, and the work efficiency is low. In addition, it does not have the ability to distinguish whether items are garbage or not, nor does it

have the ability to treat garbage by category. In fact, according to the shape, material, and other attributes of the item itself, as well as the relationship with other items, such as its location, you can further determine whether it is garbage, improve its intelligence, and avoid waste of resources; and different types of garbage should be sorted by category to meet environmental protection requirements.

In order to solve the above problems, a feasible solution is to perform intelligent garbage classification tasks on the home service robot. On the one hand, the home service robot is equipped with visual sensors to enable it to obtain visual perception capabilities [3]; on the other hand, research on effective perception detection algorithms aims to achieve the purpose of visual scene understanding and ultimately guide home service robots to autonomously perform intelligent garbage classification, improve work efficiency, and reduce energy consumption. At present, there have not been public reports about the work carried out on the autonomous garbage detection and classification of household service robots. Therefore, the realization of garbage classification and detection algorithms on household service robots has certain practical significance. However, only using the detection and classification model can only realize the identification and positioning of the garbage, and the degree of intelligence is not high. To make the robot achieve the ability of cognition and discrimination of objects in the home environment like humans, for example, humans can understand what they see, the items in the scene can be associated and imagined based on these items, not only relying on the appearance and geometric characteristics of the items, but also relying on the guidance and reasoning of the high-level prior knowledge of the items.

If you want service robots to have the ability to recognize and discriminate objects in the scene like humans, perhaps visual scene understanding can be competent. Visual scene understanding needs to understand not only the information of each entity object in the image, but also the relationship between the entity objects. Visual scene understanding, called image semantic description, is a hot issue combining machine vision and natural language processing [4–6]. Home environment information has the characteristics of diversity, semantics, and relevance. Intelligent decision-making for garbage classification based on big data of items in the scene is the key issue studied in this paper. In order to achieve the intelligent decision of whether items are garbage in the home environment, this paper proposes a garbage detection and classification method based on visual scene understanding. The main contributions of this paper are as follows: first, the construction of the scene multimodal knowledge graph. Aiming at the problem of rich and diverse semantics of items in the home environment, which is difficult to model, the knowledge graph is used to uniformly represent and store the input multimodal information; the second is to propose a garbage classification and detection model YOLOv5-Attention-KG based on visual scene understanding. Combining the improved YOLOv5m detection algorithm with the knowledge graph and deploying it to the edge device home service robot, the system has the ability to associate similar to people, which is the key to improving the system's intelligent garbage classification.

The subsequent chapters of this paper are arranged as follows. Section 2 starts with object detection, including traditional methods and deep learning, and then leads to knowledge graphs and finally mentions edge computing as a demonstration application of the system; the interrelationship between them and how to integrate them into the method proposed in this paper is shown in Section 3; Section 4 discusses the relevant analysis and verification of the experimental results of the model proposed in this paper; Section 5 summarizes the research work of this paper and the prospects for the next research work.

2. Related Work

Most of the traditional object detection algorithms are based on manual design and extraction of features, combined with the construction of classifiers, which have disadvantages such as relatively complex models and poor robustness. Deep learning is an important breakthrough in the field of artificial intelligence in the past decade. Since the multilayer convolutional neural network can autonomously extract and filter the features of different layers, compared with the traditional target detection method, the detection effect is more accurate and the generalization ability is stronger. At present, the target detection methods based on deep learning are divided into two types: one-stage and two-stage [7, 8]. Typical two-stage methods include region-based convolutional neural network (RCNN) method [9], Fast RCNN method [10], Faster RCNN method [11], region-based complete convolutional network (R-FCN) method [12], and other improvement methods [13]. A typical representative of the one-stage method is the YOLOv1 algorithm proposed by Redmon et al. [14]. Since YOLOv1 directly fits the position coordinates and confidence level, there are obvious defects. On the basis of YOLOv1, Redmon et al. proposed YOLOv2 [15], which uses the new basic network structure DarkNet-19, to delete the full connection layer and the last pooled layer, and uses the anchor frame to predict the boundary box. YOLOv3 [16] is the last version of the YOLO method proposed by Redmon et al. Two years later, Alexey et al. proposed the YOLOv4 [17] method. In order to improve the model's ability to detect accuracy and small targets, they proposed a better basic network, DarkNet-53, and used some techniques to improve performance. Another type of single-stage detection method is represented by SSD [18]. In recent years, many researchers have used deep learning technology in the research of garbage classification. Yang et al. [19] created a garbage classification data set and established a garbage classification model using support vector machines and convolutional neural networks. Mao et al. [20] used genetic algorithms to optimize the parameters of the fully connected layer of the DenseNet121 network and trained a garbage classification model with a classification accuracy of 99.60%. Literature [21] uses a self-encoding network to reconstruct the garbage classification data set and uses CNN to automatically extract features from the data set. Zhang et al. [22] used the Faster RCNN algorithm to detect

681 street pictures with 9 categories of garbage targets, and the detection mAP was 0.82, but there was a problem of unbalanced categories. Seredkin et al. [23] used Faster RCNN network to perform garbage classification which has high accuracy and effectively realized garbage identification. Chen et al. [24] used the Faster RCNN algorithm to detect 199 garbage targets on the pipeline and obtained a system missed identification rate of 3% and a false identification rate of 9%. Abeywickrama et al. [25] regarded garbage classification as image classification, used support vector machines and convolutional neural networks to identify and classify 6 types of garbage, and achieved a recognition result with a recognition accuracy of 83%. Mikami et al. [26] produced a dataset consisting of 2561 garbage images and designed a GarbNet model with an accuracy rate of 87.69%. With the increasing demand for garbage detection and classification of mobile edge devices, most of the hardware used in these scenarios is edge devices with weak computing power, and some larger detection networks are difficult to deploy. YOLOv5 launched by Ultralytics in 2020 has the advantages of small size, fast speed, and high precision and is suitable for deployment on edge devices. Therefore, this paper uses YOLOv5 as the basic network. In addition, since the above studies are based on the premise that the object is garbage, it mainly relies on a large amount of labeled data to fit a large number of parameters for prediction and lacks the guidance of prior knowledge, so the degree of system intelligence needs to be further improved. Therefore, this paper intends to add a knowledge graph on the basis of the YOLOv5 algorithm to further enhance the intelligent level of the system.

The knowledge graph aims to describe the entities, attributes, and their relationships that exist in the real world. It is generally expressed in the form of triples, so it is an effective method to use the knowledge graph to store and represent the attribute information and interrelated information of the item itself. The multimodal knowledge graph enriches the information types in the knowledge graph by combining the semantic information in the triples and the image feature information in the image and improves the information density and is widely used in question answering systems [27], search and recommendation systems [28, 29], and other fields. Literature [30] uses YOLO9000 as the object recognition module, which can recognize 9000 object categories after training, and uses external knowledge graph to obtain background knowledge related to the object. Marino et al. [31] studied the application of structured prior knowledge in the form of knowledge graph in image classification. Liu et al. [32] proposed a collection of three knowledge graphs of MMKG (Multimodal Knowledge Graphs), including the digital features and images of all entities and the overall alignment between the knowledge graphs. Chen et al. [33] proposed an expression learning framework for knowledge embedding. The framework first builds a knowledge graph based on statistical "category-attribute" related information; then it uses a graph network to spread node information on the graph to learn its knowledge expression; finally it designs a gated network to embed the knowledge expression into the image feature learning process and guides the learning of the attributes associated with the feature. Jiang et al. [34] proposed a hybrid knowledge routing module to improve model performance. In order to solve the traditional methods that ignore the correlation between the training set and the test set category, Wang et al. [35] proposed the use of category semantic expression and knowledge graph to guide the information dissemination between categories and applied it to zero-sample learning. Chen et al. [36] introduced statistical target objects and the possible coexistence of prior knowledge to constrain the relationship prediction space, aiming at improving the accuracy of the model in less sample categories. Wang et al. [37] introduced the prior knowledge of the association between the characters in the scene and the surrounding objects and performed explicit reasoning based on knowledge. Wu et al. [38] proposed a visual question-and-answer method, which constructs a textual representation of the semantic content of an image and merges it with the textual information from the knowledge base, aiming at having a deeper understanding of the scene. Lu et al. [39] combined visual features and prior knowledge of language models to determine visual relationships and realized the detection of multiple visual relationships in a picture. For object attributes such as shape and color, Sun et al. [40] proposed a method to automatically extract visual concepts using similar text and visual collections.

In order to test the effectiveness of the method proposed in this paper, consider deploying the model to edge devices for experimental verification. To perform big data analysis and management in complex systems, edge computing, as a new paradigm, can sink cloud computing functions and services to network edge devices and provide real-time data analysis and intelligent processing nearby, thereby effectively solving the problems of network congestion and network delay caused by the transmission and processing of massive data. Different from the large-scale data processing center in cloud computing, the communication, computing, storage, and other resources of edge devices in mobile edge computing are relatively limited [41]. On the one hand, when the task demands of end users increase sharply, a large number of end users need to offload tasks to edge devices, which is prone to problems such as excessive task load and increased processing delay, resulting in the lack of timeliness of task processing; on the other hand, there is an unbalanced load distribution among devices, and it is prone to the problem that some edge devices are overloaded with tasks and other edge device resources are idle. To effectively cope with the above problems, multiple edge devices can coordinate to perform computing tasks to achieve load balancing among edge devices while ensuring the service requirements of end users. Therefore, multiedge device collaboration has become an inevitable trend. The latest research work considers the collaboration of multiple edge devices to perform computing tasks together. Literature [42, 43] uses matching strategies to formulate task offloading strategies among multiple end users and multiple edge devices. Literature [44] studies the problem of task offloading in dense deployment scenarios of edge devices. Through the alliance game theory among multiple edge devices, a cooperative alliance is formed to jointly perform the computing tasks of the end user.

Literature [45, 46] implements task offloading between edge devices through a distributed game method, with the goal of minimizing the overall execution delay of the task.

3. The Design of Garbage Sorting Model

3.1. System Architecture. This paper designs a complex system for garbage detection and classification based on visual scene understanding. The overall architecture of the system is shown in Figure 1. First of all, through the knowledge graph, the unified representation and storage of the multimodal item knowledge in the home environment is used to form a priori knowledge base; among them, the YOLOv5m-Attention detection algorithm recognizes and locates the two modalities of the item image and video in the scene to obtain the item entity category and location information and combines the prior knowledge base to form a visual scene understanding model YOLOv5m-Attention-KG (see Figure 2); secondly, cloud computing is used as a computing back-end to form collaborative computing with edge devices. Finally, home service robots are used as edge computing devices for experimental verification, supporting real-time data processing and analysis, and completing the task of garbage classification.

3.2. The Key Technology and Algorithm. In order to autonomously realize intelligent garbage classification on home service robots, this paper proposes a YOLOv5m-Attention-KG visual scene understanding model. The structure of the model is shown in Figure 2. First, according to the different modalities of the items in the home environment, different model processing is adopted, and the YOLOv5m-Attention detection algorithm is used to process the two modalities of video and image; use BLSTM-LCRF and PCNN-BLSTM-Attention proposed by Wang et al. [47] to extract entities and relationships from text modalities. The open source structured data collected from the Internet and the entity relationship extracted above form a knowledge triple. The knowledge graph finally constitutes a unified characterization and storage of the semantic description information, attribute information, and spatial location information of the items in the scene. The open source structured data collected from the Internet completes the extraction of item attributes and relationship information; then it forms a knowledge triple with the entity relationships extracted above; the final knowledge graph can uniformly represent and store the semantic description, attributes, and related information of the items in the scene. Secondly, when detecting and classifying items in the home environment, the YOLOv5m-Attention detection algorithm will perform real-time detection to obtain its location and category information and query the entity information with high semantic similarity to the category information in the knowledge graph, based on the returned attributes and related information to determine whether the item is garbage and what kind of garbage it is, and make further intelligent decisions.

3.2.1. Multimodal Knowledge Graph. With the continuous popularization of the Internet technology, information from different sources such as text, images, video, and audio jointly portrays the same or related content, presents complex, multilevel semantic relationships, and forms multimodal information. As shown in Figure 3, the multimodal knowledge graph is divided into three parts: information representation, knowledge processing, and knowledge update. Entity extraction is generally to automatically extract a list of entities from a multimodal sample. At present, there is no special study on the extraction method of multimodal attribute extraction. Generally, attributes are regarded as a kind of entity concept, and the same method is used as entity extraction. Relations in multimodal samples are divided into simultaneous relations and hierarchical relations. Generally, when extracting relations, the idea that general concepts appear more frequently than specific concepts is used to extract by calculating the statistical relationship between the text and image features of the entity. Knowledge inference of multimodal samples can use label propagation based on multimodal features. For example, Fang et al. [48] use similarity matrix and image similarity matrix for label propagation; factor graphs can also be used for derivation and learning. Because every step of the construction process of the multimodal knowledge graph requires all multimodal samples, if new samples are added, a comprehensive update is required. However, there are currently no more relevant papers on the multimodal knowledge graph. The knowledge graph contains a large amount of factual knowledge, which is generally represented by triples: $(h, r, t)h$ represents the head entity, t represents the tail entity, and r represents the relationship between the two entities. The input multimodal information knowledge is modeled as a collection of triples. In the knowledge graph, nodes are used to represent entities and edges are used to represent attributes or relationships. Thus, the entities and relationships in the real indoor scene can be formed into a huge picture of the semantic network. Figure 4 is a case of knowledge graph. For the same entity as a drink bottle, due to the integrity of the shape, the attributes of the material information, and the relationship with other entities, it can be judged whether it is recyclable garbage.

3.2.2. Improved YOLOv5m-Attention Algorithm Design

3.2.2.1. Network Structure. YOLOv5 is divided into 4 models, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, according to the depth of the network and the width of the feature map. In this paper, considering the accuracy and speed, the YOLOv5m network is selected as the model for item detection and classification. YOLOv5m still uses the overall layout of v3 and v4 and divides the entire network structure into four parts: Input, Backbone, Neck, and Output. The difference from the original network is that the ESA attention mechanism is added after the Cross Stage Partial Networks (CSPNet), as shown in the highlighted module in Figure 5. Input terminal: adaptively zoom the picture, adopt Mosaic data enhancement method, enrich the data, improve the recognition ability of small objects, and automatically calculate the best anchor frame value of the data set.

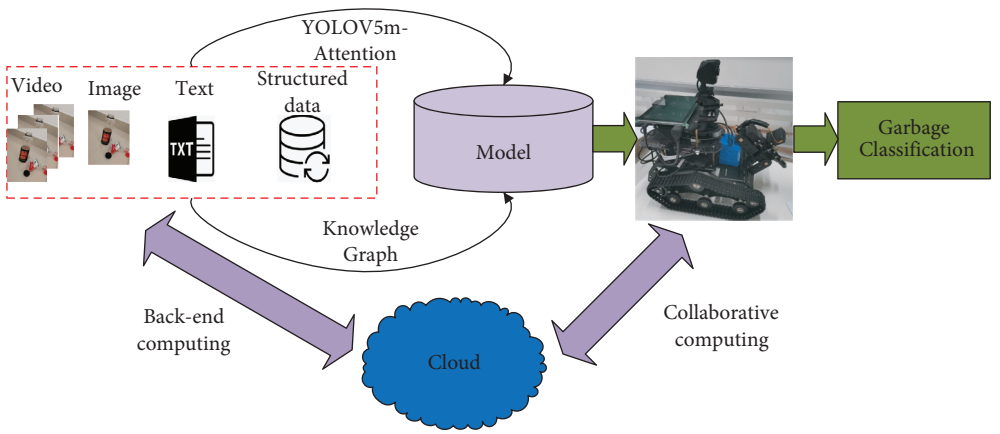

Figure 1: Overall architecture diagram.

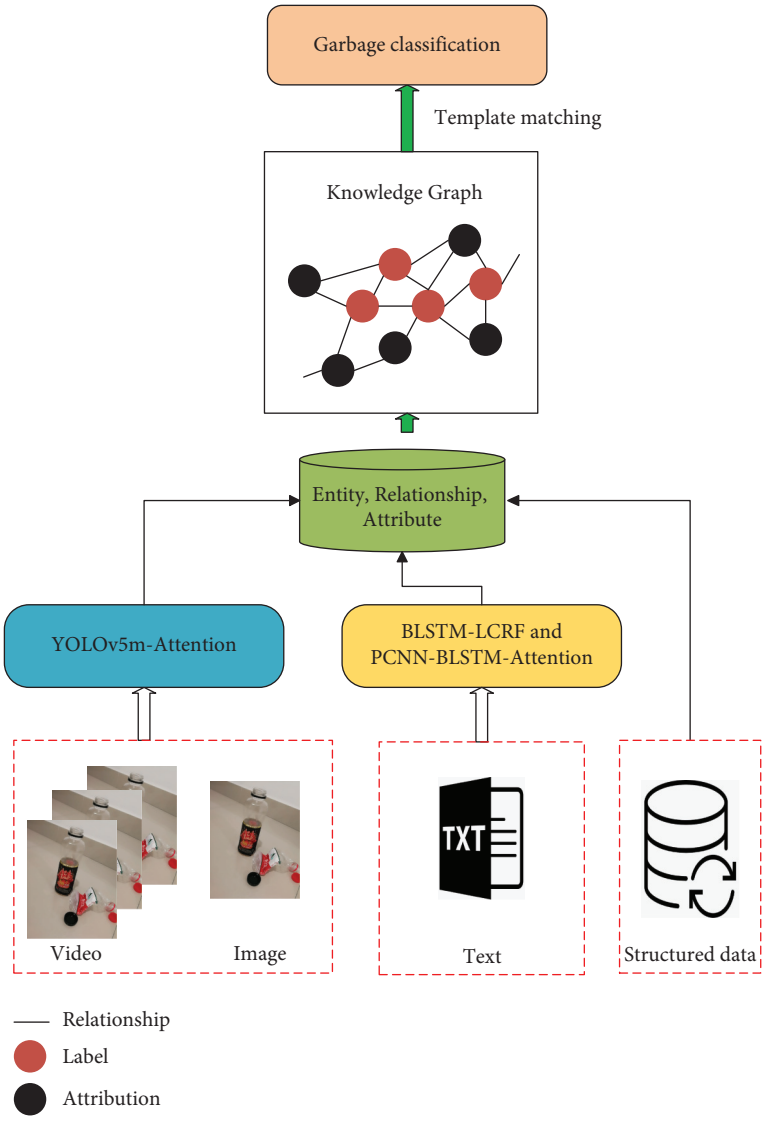

Figure 2: Visual scene understanding model.

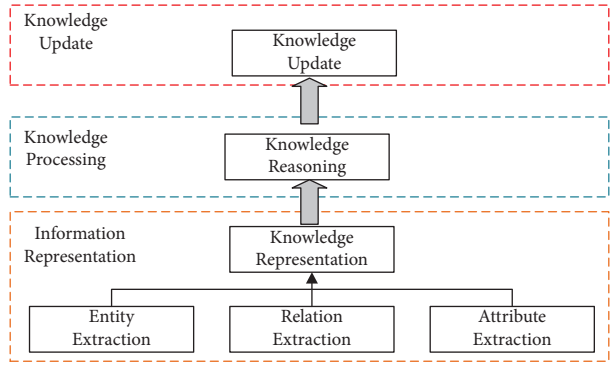

Figure 3: Construction of multimodal knowledge graph.

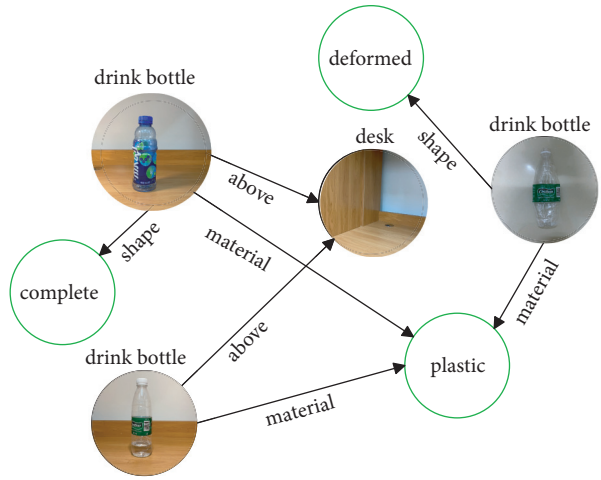

Figure 4: Knowledge graph case.

Backbone contains the Focus structure and improved CSPNet. The Focus structure includes 4 slice operations and 1 convolution operation with 32 convolution kernels, turning the original $608 \times 608 \times 3$ image into a $304 \times 304 \times 32$ feature map. CSPNet imitates the idea of dense cross-layer connection of Densenet, performs partial cross-layer fusion, and uses the feature information of different layers to obtain a richer feature map. In the figure, $n = 1$ or 2, X takes 1 or 3, which represents X residual components Res unit, a total of $n * X$ residual components Res unit. The ESA module (see Figure 6) calculates the weight information of the feature map on the channel position and spatial position and makes the network focus on the feature regions that are beneficial to classification according to the weight distribution and suppresses the background and other secondary information. Neck contains Path Aggregation Network (PANet) and Space Pyramid Pooling (SPP) modules. PANet aggregates high-level feature information with the output features of different layers of CSP modules from top to bottom and then aggregates shallow features through a bottom-up path aggregation structure, thereby fully fusing image features of different layers. The SPP module first uses 4 cores of different sizes to perform the maximum pooling operation and then performs tensor splicing. Output layer: In this paper, between GIOU Loss [49] and CIOU Loss [50], CIOU Loss with a slightly better effect is finally selected as the loss function of the prediction box regression. Because CIOU Loss considers the scale information of the bounding box aspect ratio compared to GIOU Loss and measures it from the three angles of overlap area, center point distance, and aspect ratio, it makes the prediction box regression better.

Drawing lessons from the ideas of CBAM [51] and ECA attention mechanism [52], the ESA attention block first obtains the channel and spatial attention weight maps according to the input feature map of the model; then it, respectively, multiplies it with the original feature map to obtain the space and channel feature maps with weights; finally, the channel and space feature maps are added in parallel to obtain a feature map with attention weights. The ESA attention structure is shown in Figure 6.

The difference from the CBAM attention mechanism is that the channel attention CAM in the ESA attention mechanism borrows the ECA attention mechanism. After global average pooling of the input features, it does not change the dimension of the channel and uses the size k Fast one-dimensional convolution to capture the local cross-channel feature information of each channel, replacing the multilayer MLP block in the channel attention mechanism in CBAM, avoiding the problem of reduced attention to the channel caused by dimensionality

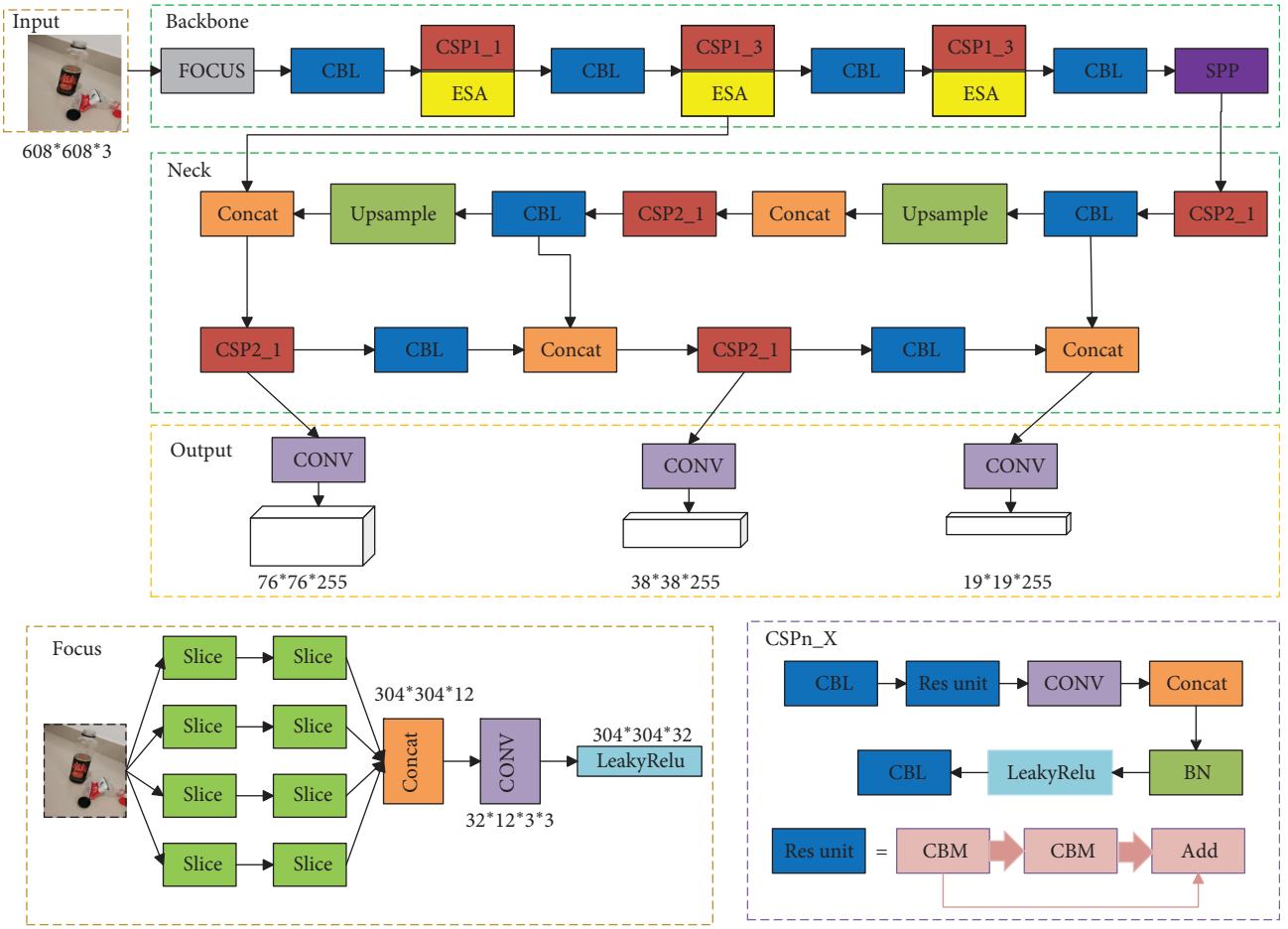

Figure 5: YOLOv5m-Attention network structure diagram.

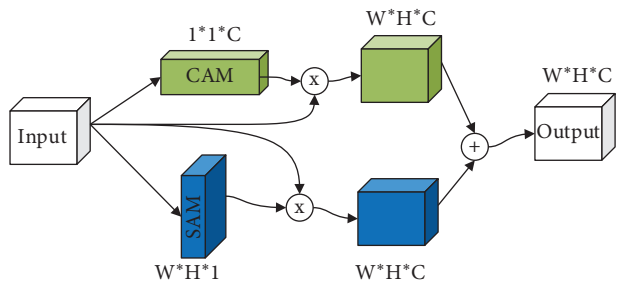

Figure 6: ESA attention structure.

reduction in MLP, and at the same time significantly reducing the complexity of the model. The spatial attention SAM in the ESA attention mechanism performs global average pooling and maximum pooling operations on the pixel values of the same position on the input feature map at the spatial position and obtains two spatial attention weights, respectively. They are merged into a 2-channel feature map in the channel dimension. Then use the convolutional layer composed of $3 * 3$ convolution kernel to compress the channel to 1, get the feature map size $W * H * 1$, and finally activate it through the Sigmoid function to obtain the spatial attention block.

3.2.2.2 Loss function. The model loss function is composed of classification loss, localization loss, and object confidence loss. YOLOv5 uses binary cross entropy loss to calculate the loss of category probability and object confidence score. Through experiments, the loss function CIOU_Loss is shown in formula (1).

Input: Self-made object image data set and markup files
Initialization parameters: epoch, learning rate, batch size, input size, network model configuration yaml file, IOU, yolov5m.pt pre-training weights
Image preprocessing: image brightness, image contrast, image saturation, Mosaic
(1) Prepare data, make data set, and divide training set and validation set.
(2) Load data configuration information and initialization parameters, input data, and preprocess it.
(3) Load the network model, and perform feature extraction and object positioning and classification on the input image.
(4) As the number of epochs increases, use SGD to update and optimize each set of parameters in the network.
(5) If the current epoch is not the last round, the MAP of the current model is calculated in the validation set. If the calculated model performance is better, the best model is updated and stored.
(6) After training the set number of epochs, obtain the trained optimal performance model and the most recently trained model.
Output: The best-performing detection model in this training.

Algorithm 1: Model training algorithm.

$$\text{CIOU\_Loss} = 1 - \left\{\text{IOU} - \frac{d_1^2}{d_2^2} - \beta\alpha\right\}. \tag{1}$$

Among $\beta = (\alpha/(1 - IOU) + \alpha)$, $\alpha = (4/\pi^2) = (\tan^{-1}(W^{gt}/h^{gt}) - \tan^{-1}(W/h))^2$, d_1 represents the Euclidean distance between the two center points of the prediction box and the object box, and $d_2$represents the diagonal distance of the smallest bounding rectangle. (W^{gt}/h^{gt}) and (W/h), respectively, represent the respective aspect ratios of the object frame and the prediction frame.

3.2.2.3 Network training. The overall training process of the YOLOv5m network (Algorithm 1):

Some network parameter descriptions are shown in Table 1.

Table 1: Network parameter description table.

Parameter name	Parameter values
Learning rate	0.001
Momentum	0.9
Decay	0.0005
Batch size	64

4. Experiment

4.1. Experimental Configuration. The experiment in this paper is built on the Windows environment. CUDA is a general parallel computing architecture launched by NVIDIA. CUDNN is a GPU acceleration library for deep neural networks. The data is trained through the cooperation of the two. The experimental configuration is shown in Table 2.

4.2. Data Collection. The data set used in this paper has a total of 15,000 domestic garbage pictures, most of which come from the data set in the garbage classification competition held by Ali Yun Tianchi and some pictures of domestic garbage collected by the author. The data set can be divided into four categories in general, namely, recyclable trash, food trash, harmful trash, and other trash. Each category contains multiple objects. Among them are recyclable trash: power bank, bag, wash supplies, plastic toy, plastic utensils, plastic hangers, glassware, metalware, courier bags, plug wire, old clothes, ring-pull can, pillow, plush toy, shoes, cutting board, carton, wine bottle, metal food can, ironware, wok, edible oil drum, drink bottle, and paper books; harmful trash: dry battery, Unguentum, and expired drugs; other trash: disposable snack box, stained plastic, but, toothpick, flowerpot, chinaware, chopsticks, and stained paper; 10% of each category selects a total of 1500 images as the validation set, and the remaining 13,500 images are used as the training set. Use the LabelImage tool to label the training set, and generate the corresponding xml file for training. Figure 7 shows a visual display of the data set. Figure 7 shows a visual display of the data set. The left picture is the label distribution map of the data set. The sample distribution of various items can be clearly seen. There are many samples of small and large targets; the right picture is the distribution of data correlation maps.

4.3. Experimental Indicators. The evaluation criteria of the results of this experiment are mainly Precision (P), Recall (R), Mean Average Precision (MAP), and detection speed FPS. Among them, Precision represents the ratio of the real samples in the recognized positive samples.

P represents the ratio of the total number of predicted correct positive samples to the total number of actual positive samples in the prediction data set, as shown in formula (2):

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

where R represents the probability that the correct category in the sample is predicted to be correct, as shown in formula (3):

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

The MAP is determined by the precision rate P and the recall rate R. The curve with R as the horizontal axis and P as the vertical axis is referred to as the PR curve. The area under the PR curve is recorded as the AP value, as shown in formula (4), and the average of the average accuracy of all object categories is the MAP value, as shown in formula (5):

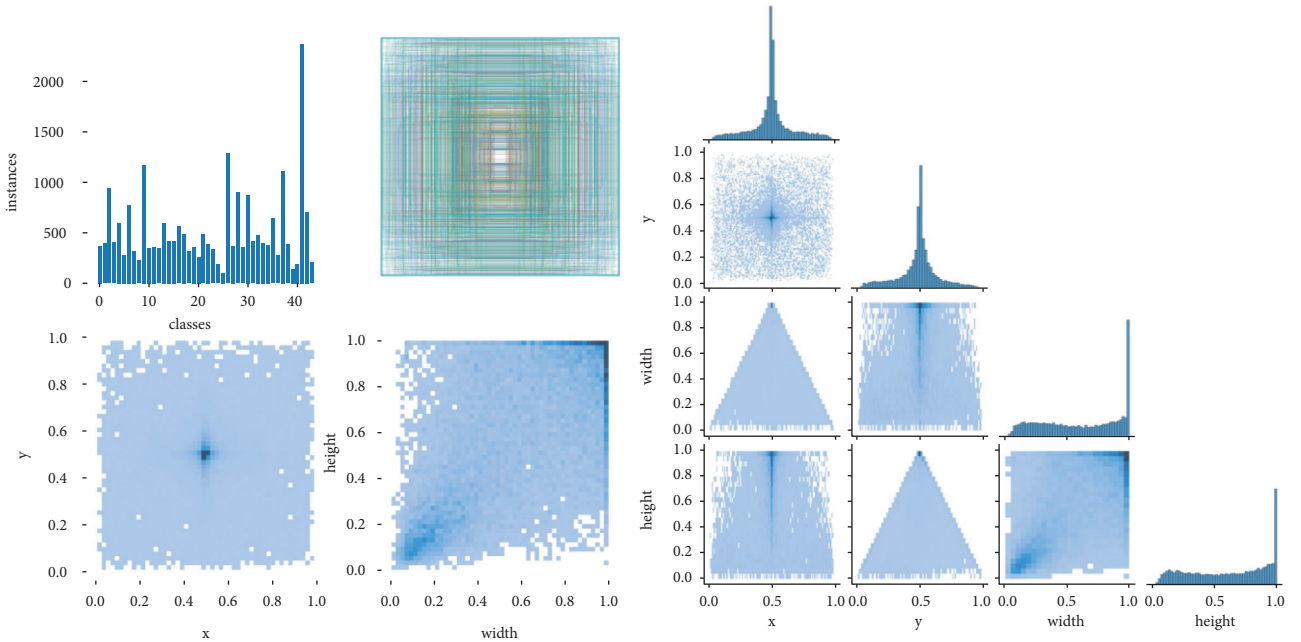

Figure 7: Data set visualization.

$$\mathrm{AP} = \int_0^1 P \mathrm{d}R, \tag{4}$$

$$\mathrm{MAP} = \frac{\sum_{i=1}^{N} \mathrm{AP}_i}{N}. \tag{5}$$

Among them, TP in formulae (2) and (3) indicates that the correct class is predicted as the number of correct categories, FP indicates that the negative category is predicted as the number of correct categories, and FN indicates that the correct category is predicted as the number of negative categories; in formula (5) N refers to the total number of detection object categories.

Table 2: Experiment environment configuration.

Project	Experimental environment
System	Windows
Programming environment	Pycharm
GPU	NVIDIATITAN RTX
Memory	24 GB
Pytorch version	Pytorch1.6
Python version	Python3.6
CUDA version	CUDA10.1
CUDNN version	CUDNN7.6
Data bases	Neo4j4.2.2
Java runtime environment	JDK15.0.1

4.4. Experimental Results and Analysis. The network parameters were trained and verified with different algorithms in accordance with Table 2 above, and the MAP and FPS were calculated as shown in Table 3. Figure 8 shows part of the evaluation indicators of the improved model in turn. The upper left is the training and verification loss function curve. You can intuitively see that, after 100 epochs, the loss reaches the lowest value and tends to balance; the upper right is the confusion matrix; the lower left is the PR curve; the bottom right is the $F1$ value curve.

From the results in Table 3, compared with the original YOLOv5 algorithm, the average accuracy rate of YOLOv5m-Attention-KG is increased by about 0.4% when the detection speed is equivalent. It also shows that the algorithm has a lower cost of additional propagation time in exchange for detection accuracy. Figure 9 is a partial visual comparison result of the original YOLOv5 and YOLOv5m-Attention-KG algorithms, where Figures 9(a)–9(k) are the detection results of the original YOLOv5 algorithm and Figures 9(b)–9(l) are the corresponding detection results of the YOLOv5m-Attention-KG algorithm. With improved algorithm from the comparison of Figures 9(a)–9(g) and 9(b)–9(h), it is obvious that the accuracy rate has improved; Figures 9(i)–9(k) and Figures 9(j)–9(l) show YOLOv5m-Attention-KG algorithm to increase the missed detection rate, and the accuracy rate has been improved.

Figure 10 shows the application of garbage classification. It is the same as the entity label of the drink bottle, because it has different attributes and the position relationship with other entity labels can make different decisions. The two entities of the drink bottle and the desk can obtain their entity labels through the recognition algorithm, and the entity label is used as a keyword to query in the neo4j graph database, and an intelligent decision will be made as to whether it is garbage. For example, the drink bottle in Figure 10(a) is placed on the floor, and because its shape is deformed and the material is plastic, it can be concluded that it is recyclable garbage, while the drink bottle in Figure 10(b) is placed on the desk and its shape is complete, so it cannot be determined to be garbage.

Table 3: Parameters of detection results of each algorithm.

Algorithm	MAP (%)	FPS
YOLOv5	72.8	30
YOLOv5m-Attention-KG	73.2	28

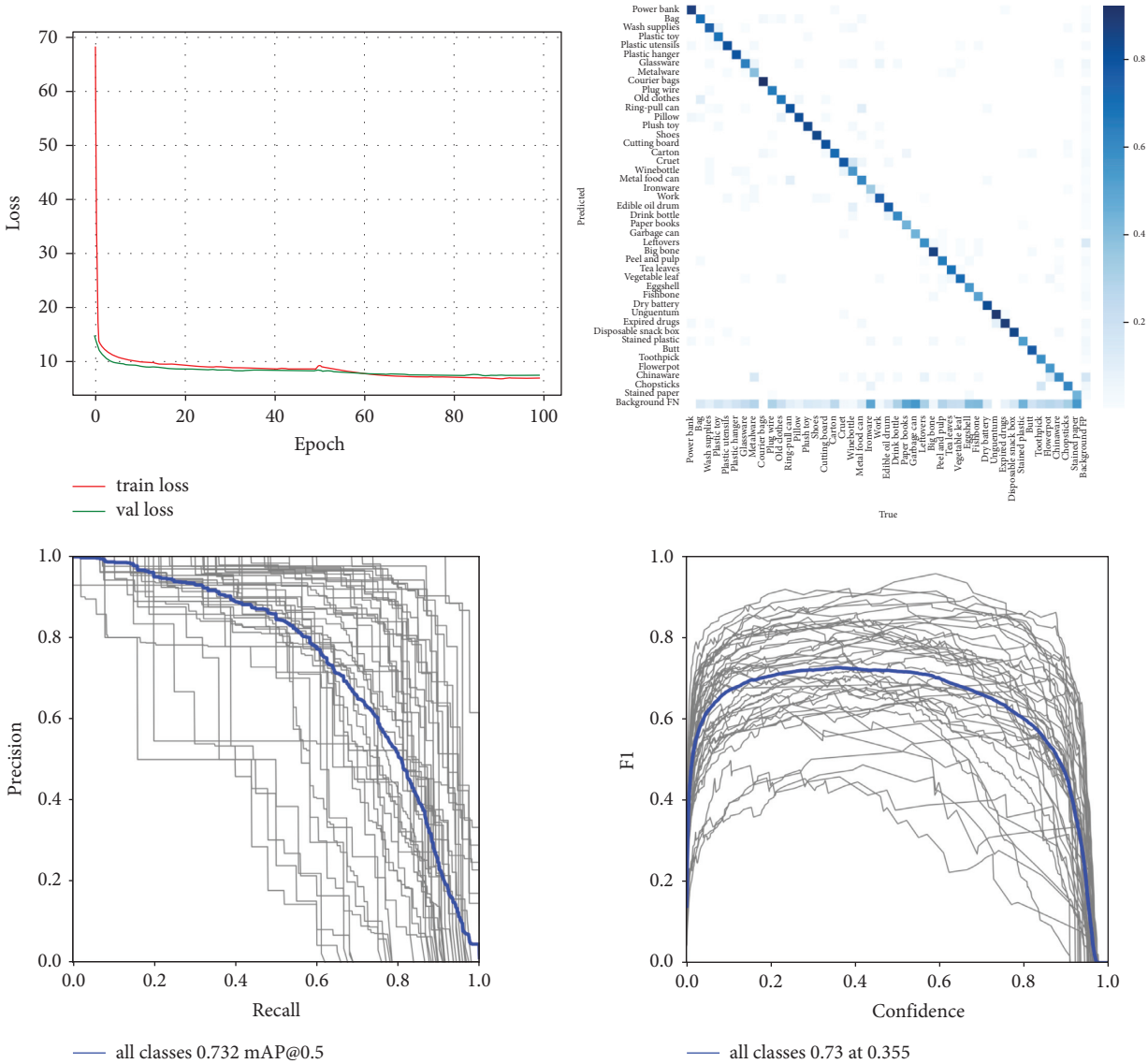

Figure 8: Part of the evaluation index.

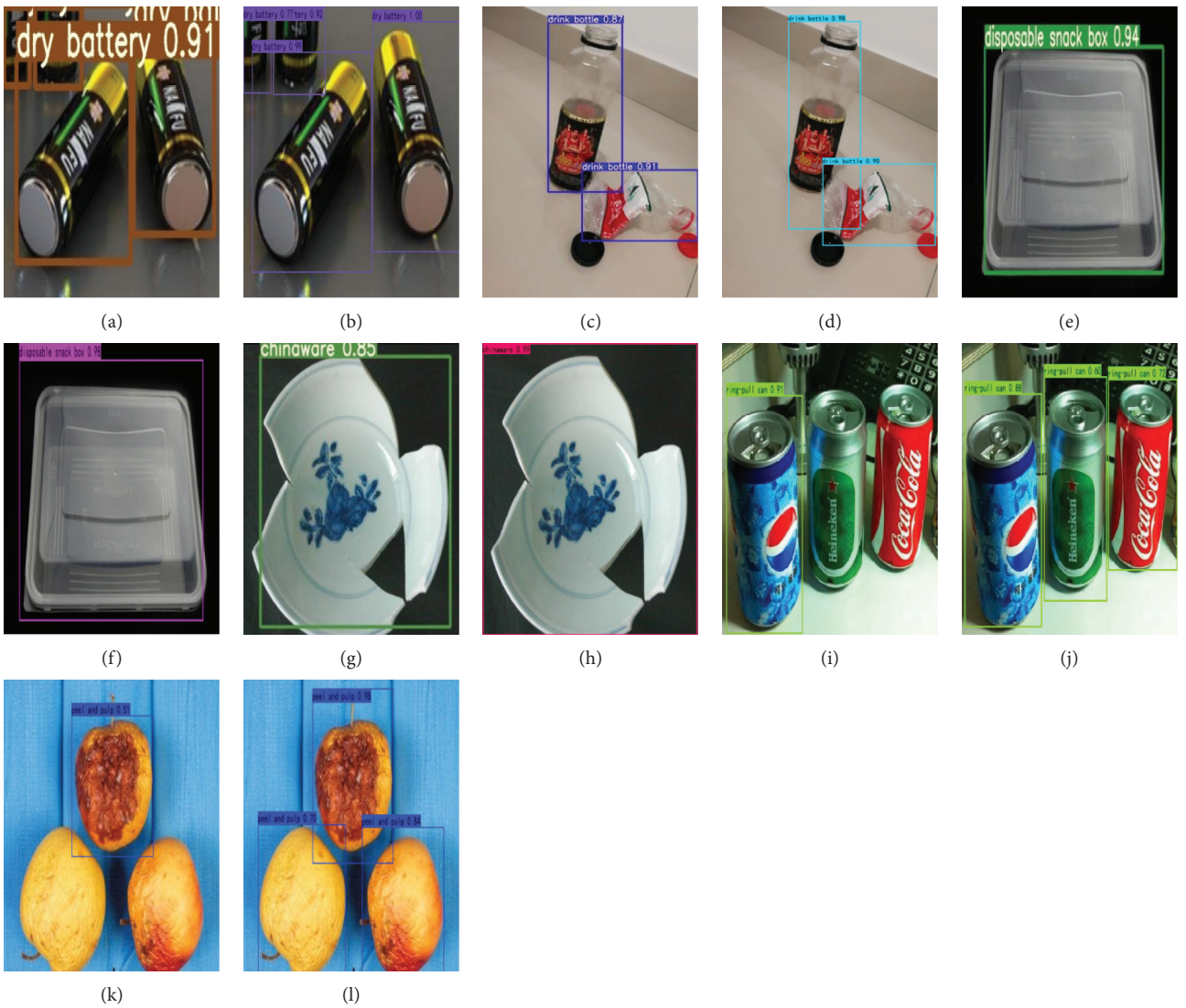

Figure 9: Comparison result graphs.

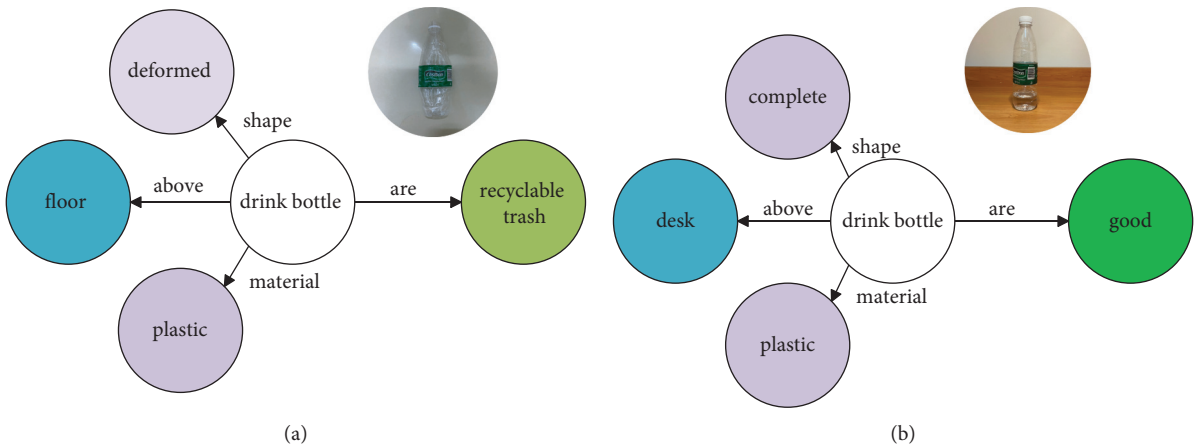

Figure 10: Garbage classification.

5. Conclusions and Future Work

In order to autonomously complete intelligent garbage classification tasks on edge devices, this paper proposes a garbage detection and classification method based on visual scene understanding. Different from the existing method, the perceptual detection under the premise that the item is artificially defaulted to be garbage, this method uses knowledge graphs and visual algorithms to realize intelligent decision-making of items in the scene. Potential future research directions: First, the extraction of the attributes of the items in the scene and the associated information of other items requires further in-depth research; the second is that the system is now only real-time perception of two modal items of image and video, and it can go deep into voice modal in the future, through intelligent interaction with people, to improve the degree of intelligence of edge devices.

Data Availability

The data used to support the findings of this study are not applicable because the data interface cannot provide external access temporarily.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by the National Key R&D Program of China (Grant nos. 2019YFE0122600, 2018YFB1700200, and 2019QY1604), National Natural Science Foundation of China (Grant no. U1836217), Hunan Provincial Key Research and Development Project of China (Grant no. 2019GK2133), Natural Science Foundation of Hunan Province (Grant nos. 2021JJ50050, 2021JJ50058, and 2020JJ6089), Scientific Research Project of Hunan Provincial Department of Education (Grant no. 19B147), Key Project of the Department of Education in Hunan Province (Grant no. 19A133), Scientific Research Project of China Packaging Federation (Grant no. 17ZBLWT001KT010), Open Platform Innovation Foundation of Hunan Provincial Education Department (Grant no. 20K046), and Special Fund Support Project for the Construction of Innovative Provinces in Hunan (2019GK4009).

References

[1] L. K. Zhou, H. Z. Liu, and Y. Li, "Summary for the key technologies and research status of the cleaning robot," *Mechanical Science and Technology for Aerospace Engineering*, vol. 33, no. 5, pp. 635–642, 2014.

[2] D. Q. Zhu and M. Z. Yan, "Survey on technology of mobile robot path planning," *Control and Decision*, vol. 25, no. 7, pp. 961–967, 2010.

[3] J. Y. Yang, L. Ma, and D. C. Bai, "Robot vision environmental perception method based on hybrid features," *Journal of Image and Graphics*, vol. 17, no. 1, pp. 114–122, 2012.

[4] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228, Salt Lake City, UT, USA, June 2018.

[5] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3242–3250, Honolulu, HI, USA, July 2017.

[6] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1151–1159, Honolulu, HI, USA, July 2017.

[7] Y. Zhang, C. Song, and D. Zhang, "Deep learning-based object detection improvement for tomato disease," *IEEE Access*, vol. 8, pp. 56607–56614, 2020.

[8] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Towards deep object detection techniques for phoneme recognition," *IEEE Access*, vol. 8, pp. 54663–54680, 2020.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[10] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[11] 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, 2016.

[12] J. Dai, Y. Li, and K. He, "R-FCN: object detection via region-based fully convolutional networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 379–387, Barcelona, Spain, December 2016.

[13] K. N. R. S. V. Prasad, K. B. D'souza, and V. K. Bhargava, "A downscaled faster-RCNN framework for signal detection and time-frequency localization in wideband RF systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4847–4862, 2020.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[15] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, Honolulu, HI, USA, July 2017.

[16] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[17] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: single shot multiBox detector," in *Proceedings of the European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[19] M. Y. Yang and G. Thung, "Classification of trash for recyclability status," *Computer Science*, pp. 1–6, 2016.

[20] W. L. Mao, W. C. Chen, C. T. Wang, and Y. H. Lin, "Recycling waste classification using optimized convolutional neural

network," *Resources, Conservation and Recycling*, vol. 164, 2021.

[21] M. Toaar, B. Ergen, and Z. Cmert, "Waste classification using AutoEncoder network with integrated feature selection method in convolutional neural network models," *Measurement*, vol. 153, 2019.

[22] P. Zhang, Q. Zhao, J. Gao, W. Li, and J. Lu, "Urban street cleanliness assessment using mobile edge computing and deep learning," *IEEE Access*, vol. 7, pp. 63550–63563, 2019.

[23] A. Seredkin, M. Tokarev, and I. Plohih, "Development of a method of detection and classification of waste objects on a conveyor for a robotic sorting system," *Journal of Physics: Conference Series*, vol. 1359, 2019.

[24] C. Zhihong, Z. Hebin, W. Yanbo, L. Binyan, and L. Yu, "A vision-based robotic grasping system using deep learning for garbage sorting," in *Proceedings of the 2017 36th Chinese Control Conference (CCC)*, pp. 11223–11226, Dalian, China, July 2017.

[25] T. Abeywickrama, M. A. Cheema, and D. Taniar, "k-nearest neighbors on road networks," *Proceedings of the VLDB Endowment*, vol. 9, no. 6, pp. 492–503, 2016.

[26] K. Mikami, Y. Chen, J. Nakazawa, Y. Iida, Y. Kishimoto, and Y. Oya, "Deep counter: using deep learning to count garbage bags," in *Proceedings of the 2018 IEEE 24th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pp. 1–10, Hakodate, Japan, August 2018.

[27] L. Bauer, Y. Wang, and M. Bansal, "Commonsense for generative multi-hop question answering tasks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4220–4230, Brussels, Belgium, 2018.

[28] H. W. Wang, M. Zhao, X. Xing, W. J. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *Proceedings of the World Wide Web Conference*, San Francisco, CA, USA, March 2019.

[29] Y. Z. Wu, Q. Liu, R. Chen, C. Y. Li, and Z. R. Peng, "A group recommendation system of network document resource based on knowledge graph and LSTM in edge computing," *Security and Communication Networks*, vol. 2020, Article ID 8843803, 11 pages, 2020.

[30] Y. Zhou, Y. W. Sun, and V. Honavar, "Improving image captioning by leveraging knowledgegraphs," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 283–293, Waikoloa Village, HI, USA, January 2019.

[31] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: using knowledge graphs for image classification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20–28, Honolulu, HI, USA, July 2017.

[32] Y. Liu, H. Li, and A. Garcia-Duran, "MMKG: multi-modal knowledge graphs," 2019, https://arxiv.org/abs/1903.05485.

[33] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, "Knowledge-embedded representation learning for fine-grained image recognition," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI*, Stockholm, Sweden, July 2018.

[34] C. Jiang, H. Xu, X. Liang, and L. Lin, "Hybrid knowledge routed modules for large-scale object detection," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1552–1563, Montréal, Canada, December 2018.

[35] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6857–6866, Salt Lake City, UT, USA, June 2018.

[36] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6156–6164, Long Beach, CA, USA, June 2019.

[37] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, and L. Lin, "Deep reasoning with knowledge graph for social relationship understanding," 2018, https://arxiv.org/abs/1807.00504.

[38] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: free-form visual question answering based on knowledge from external sources," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4622–4630, Las Vegas, NV, USA, June 2016.

[39] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proceedings of the 14th European Conference on Computer Vision*, pp. 852–869, Amsterdam, The Netherlands, October 2016.

[40] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2596–2604, Santiago, Chile, December 2015.

[41] Q. Zhang, S. Sun, M. Liu, Z. Li, and Z. Zhang, "Online joint optimization mechanism of task offloading and service caching for multi-edge device collaboration," *Journal of Computer Research and Development*, vol. 58, no. 6, pp. 1318–1339, 2021.

[42] Q. V. Pham, T. LeAnh, H. Nguyen, and C. Hong, "Decentralized computation offloading and resource allocation in heterogeneous metworks with mobile edge computing," 2018, https://arxiv.org/abs/1803.00683.

[43] X. H. Shen, Y. Z. Wu, S. H. Chen, and X. M. Luo, "An intelligent garbage sorting system based on edge computing and visual understanding of social internet of vehicles," *Mobile Information Systems*, vol. 2021, Article ID 5231092, 12 pages, 2021.

[44] L. Chen and J. Xu, "Socially trusted collaborative edge computing in ultra dense networks," in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, San Jose, CA, USA, October 2017.

[45] L. Chen, J. Xu, and S. Zhou, "Computation peer offloading in mobile edge computing with energy budgets," in *Proceedings of the 2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, December 2017.

[46] L. Chen, S. Zhou, and J. Xu, "Computation peer offloading for energy-constrained mobile edge computing in small-cell networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 4, pp. 1619–1632, 2018.

[47] H. Wang, W. Q. Zhu, Y. Z. Wu, P. J. He, and L. J. Wan, "Named entity recognition based on equipment and fault field of CNC machine tools," *Journal of Engineering Science*, vol. 42, no. 4, pp. 476–482, 2020.

[48] Q. Fang, C. Xu, J. Sang, M. S. Hossain, and A. Ghoneim, "Folksonomy-based visual ontology construction and its applications," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 702–713, 2016.

[49] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, Long Beach, CA, USA, June 2019.

[50] Z. Zheng, P. Wang, W. Liu, and D. Ren, "Distance-IoU loss: faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[51] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proceedings of the ECCV*, Munich, Germany, September 2018.

[52] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: efficient channel attention for deep convolutional neural networks," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11531–11539, Seattle, WA, USA, June 2020.