International Journal of Advanced Culture Technology Vol.9 No.3 305-314 (2021)
DOI https://doi.org/10.17703/IJACT.2021.9.3.305

IJACT 21-9-40

Municipal waste classification system design based on Faster-RCNN and YoloV4 mixed model

[1]Gan Liu, [2]Sang-Hyun Lee

[1]student., Computer Engineering, Honam University, Korea
[2]Associate Professor., Department of Computer Engineering, Honam University, Korea
1060110388@qq.com, leesang64@honam.ac.kr

Abstract

Currently, due to COVID-19, household waste has a lot of impact on the environment due to packaging of food delivery. In this paper, we design and implement Faster-RCNN, SSD, and YOLOv4 models for municipal waste detection and classification. The data set explores two types of plastics, which account for a large proportion of household waste, and the types of aluminum cans. To classify the plastic type and the aluminum can type, 1,083 aluminum can types and 1,003 plastic types were studied. In addition, in order to increase the accuracy, we compare and evaluate the loss value and the accuracy value for the detection of municipal waste classification using Faster-RCNN, SDD, and YoloV4 three models. As a final result of this paper, the average precision value of the SSD model is 99.99%, the average precision value of plastics is 97.65%, and the mAP value is 99.78%, which is the best result.

Keywords: *Faster-RCNN, SDD, YOLOv4, Municipal Waste, Region Proposal Networks*

1. INTRODUCTION

Currently, due to COVID-19, household waste has a significant impact on the environment due to the packaging of food delivery. The amount of household waste generated per day in Korea is the National Waste Generation and Treatment Status in 2019 presented by the resource circulation information system(RCIS).

Increased, and the domestic household waste recycling rate in 2019 was 59.7%, which was low [1]. However, a large amount of household waste is not sorted in various places and many problems arise in the collection and treatment of household waste and other operations. The previously used classification method of household waste is manual sorting manually or image classification by computer vision, so work intensity is high, classification efficiency is low, and work environment facilities are a bad problem arose. In order to reduce the manual waste sorting workload, deep learning-based studies of municipal waste classification and recognition methods can effectively solve these problems, and allow machines operated by artificial intelligence to do it automatically, it may be possible to detect and classify various types of waste without the need for human labor.

Recently, with the continuous development of artificial intelligence, target detection, image processing and classification, image recognition, autonomous driving, natural language processing, and recommendation systems are being used. Using the theoretical results of deep learning and machine learning in everyday life it is becoming [2].

Here, target detection of deep learning is based on target shape and statistical characteristics, and accuracy of image segmentation and real-time performance are important functions of the whole system. In particular,

Manuscript received: July 31, 2021 / revised: August 25, 2021 / accepted: August 30, 2021
Corresponding Author: leesang64@honam.ac.kr
Tel: +82-62-940-5285, Fax: +82-62-940-5285
Associate Professor, Dept. of Computer Engineering, Honam University, Korea

automatic target extraction and recognition is critical when multiple targets must be processed in real time in complex scenes. Due to advancement in computer technology and the widespread application of computer vision principles, the use of computer image processing technology to track a target in real time is becoming more and more popular. In general, dynamic real-time tracking and target, positioning has broad application value in intelligent traffic and intelligent monitoring systems, military target detection, and medical surgery, etc [3].

General target detection algorithm is divided into mainly two categories, one stage and two stage. One stage uniformly performs high-density sampling at different locations of images based on the concept of regression, extracts features using a convolutional neural network (CNN) model and performs classification and regression, which is used here algorithms include single shot multibox detector (SSD) and you only look once (YOLO) [4]. The two stage first generates a series of sparse matrix candidate frames through the algorithm network, then classifies and regresses the candidate frames to complete target detection. Algorithms applied here include faster R-CNN and Mask R-CNN.

This paper intends to implement the detection and classification of household waste using deep learning. The dataset used here is intended to explore the two types of plastic cans and aluminum cans that are most commonly used as household waste. Plastic type data 1,083 pieces and aluminum can type data 1,003 pieces were used to search and classify domestic waste. In addition, it has intended to compare and evaluate the loss value and accuracy of the results of the detection of municipal waste classification using the three models of Faster-RCNN, SDD, and YoloV4.

2. RELATED RESEARCH

In this paper, we try to compare the performance of image detection by setting up Faster-RCNN, YoloV4 model, and SSD model.

2.1 YOLOv4 Model

The network structure of YOLOv4 consists of three parts: Backbone, Neck, and Head. The backbone is composed of CSPDarknet53, the neck is composed of SPP (Spatial Pyramid Pooling) and PANet (Path Aggregation Network), and the head is composed of YOLO Head. The structure of YOLOv4 is shown in Figure 1 [5].

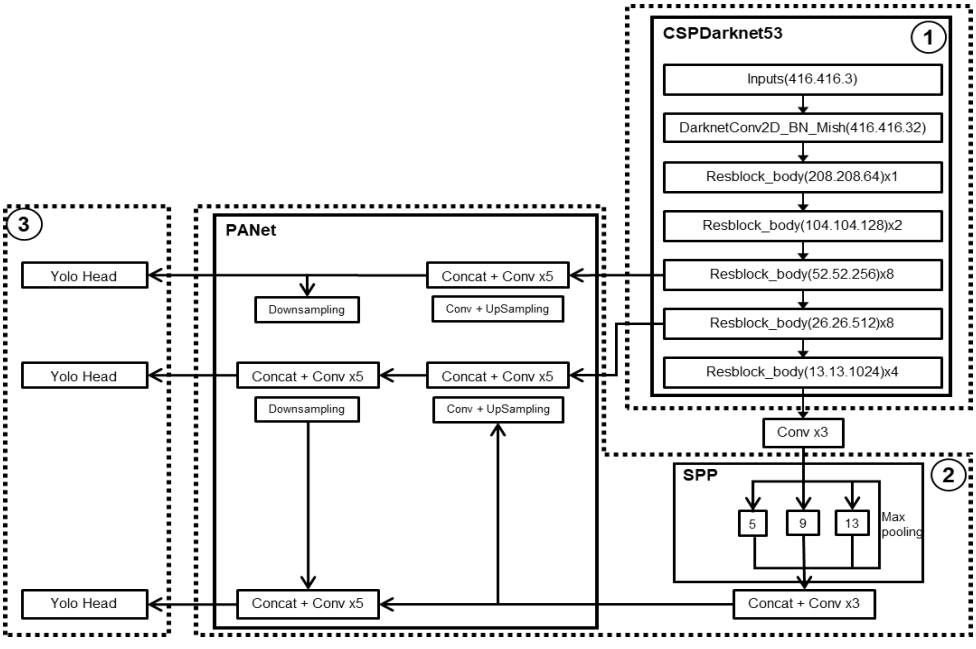

Figure 1. Structure of YOLOv4

As an explanation for Figure 1, CSPDarknet53 of YOLOv4 structure ①adds Cross Stage Partial (CSP) in Darknet-53 structure of YOLOv3 to improve the learning ability and maintain accuracy of CNN and reduce memory consumption can be reduced.

Image input to CSPDarknet53 first performs DarkNet convolution to get feature maps of 416 * 416 * 32, and then continues with five residual networks. Down sampling is also performed for 208 * 208*64, 104 * 104*128, 52 * 52*256, 26 * 26*512, and 13 * 13*1024 feature maps. The purpose of using Spatial Pyramid Pooling (SPP) in ②YOLOv4 is to improve the network coverage area. Here, run pooling(1 * 1, 5 * 5, 9 * 9, 13 * 13) for network layer entered into SPP. PANet first performs down sampling and up sampling and performs fusion of multiple features to extract and output three effective feature layers 76 * 76, 38 * 38, 19 * 19 of the network. Finally, YOLO head of ③can analyze the parameters for the type and prediction of the output feature image.

2.2 SSD Model

The Single Shot Multibox Detector (SSD) model builds a new network structure based on the Visual Geometry Group (VGG) network model [6]. Fusing the feature maps of various convolutional layers to improve the feature representation ability of the network and adopt a multi-scales convolution detection method. Performs target detection to greatly speed up target detection. The network structure of SSD is shown in Figure 2 [7].

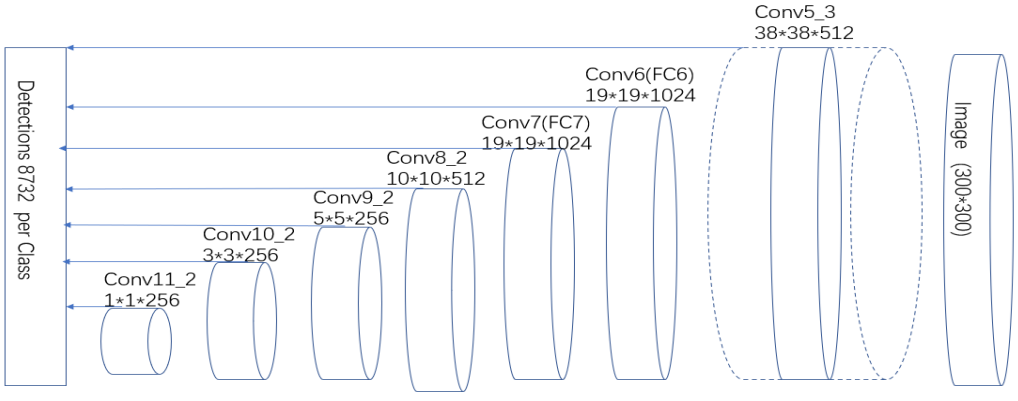

Figure 2. Structure of Visual Geometry Group Model network of SSD

The network of the SSD model is divided into two parts. First, with the network of the adjusted VGG-16 model, dropout, FC8 and softmax layers are removed, and FC6 and FC7 are replaced with Conv6 and Conv7 convolutional layers. Second, as a feature detection network behind, the four convolution layer groups Conv8, Conv9, Conv10, and Conv11 are added together, and Conv5 and Conv7 are added together to form a multi-scales feature pyramid structure convolutional network [6].

Perform feature extraction on each convolutional layer to get feature maps of different scales, and finally send six layers of feature maps of different scales to a classification and regression network to perform regression predictions for location and type of subject.

2.3 Faster R-CNN Model

Faster R-CNN is an improvement based on Faster R-CNN, proposed by Shaoqing Ren in 2017 [8]. Faster R-CNN combines feature extraction, candidate region generation, bounding box regression and object

classification into one network. The detection speed is accelerated to some extent under conditions that guarantee accuracy. The basic structure of Faster R-CNN is shown in Figure 3 [9].

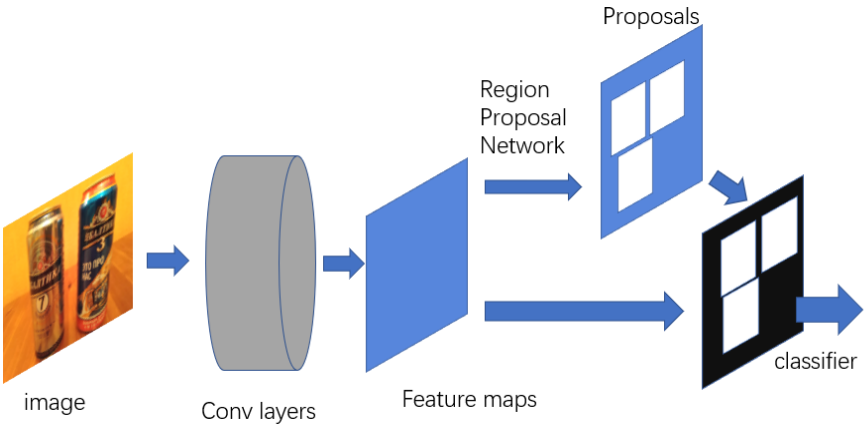

Figure 3. Structure of Visual Geometry Group Model network of Faster R-CNN

The description of the Faster R-CNN structure in Figure 2 proceeds as follows. First, conv(Convolution) layers are CNN network target detection method, Faster RCNN first extracts feature maps using a set of basic conv + relu + pooling layers. Feature maps are shared for subsequent Region Proposal Networks (RPN) layers and fully connected layers.

Region Proposal Networks (RPN) generate region proposals, check whether the anchors are positive or negative via softmax, and then correct the anchors using bounding box regression to collect accurate proposals values. The Roi Pooling layer collects the input feature maps and proposals, combines the two pieces of information, extracts the proposal feature maps, and sends them to the subsequent fully connected layer to determine the target type. Classification uses proposal feature maps to compute the types of proposals and at the same time uses bounding box regression to obtain the final precise location of the detection frame.

3. IMPLEMENTATION

3.1 Data set

According to Korea's Dobong-gu Cleaning Administration website in Seoul, Korea, the types of household waste that can be recycled are paper, vinyl, (iron) cans, Styrofoam, plastics (bottles), cartons (cups), glass (bottles), etc [10]. The study of this paper created a dataset by selecting two types of plastic and can (iron) types, which are mainly found in convenience stores and marts, out of seven types of recycling waste. Among the 2,086 total data images, 1,003 cans of beverages or alcoholic beverages were collected as can types, and 1,083 images of beverages/alcohol bottles and detergent bottles were collected for plastic types to create a Dataset. The classification and items of the data set are shown in Table 1.

Table 1. Data set type

Category	Item	amount
Aluminum can (iron)	grocery cans	540
	Beverage/Alcoholic Cans	543
plastic (bottle)	Beverage/Liquor Bottles	500
	detergent bottle	503
	Total	2,086

3.2 According Design of the proposed models

In this paper, we set up Faster-RCNN, YOLOv4 model and SSD model to compare the performance of image detection accuracy of two municipal wastes.

3.2.1 YOLOv4 model

YOLOv4 is a one-stage algorithm with fast detection speed, which can be applied more easily to engineering practice. YOLOv4 uses Mosaic data augmentation Mosaic data augmentation is a method proposed in YOLOv4 based on CutMix data augmentation [11]. Mosaic data augmentation combines four training images into one with a specific ratio. CutMix data augmentation combines images by cutting a portion from one image and pasting it into the enlarged image. The SPP [12] (Spatial Pyramid Pooling) module is added to strengthen the backbone function to largely separate the most important context functions. The model of YOLOv4 is shown in Figure 4, and the description of the image is described below.

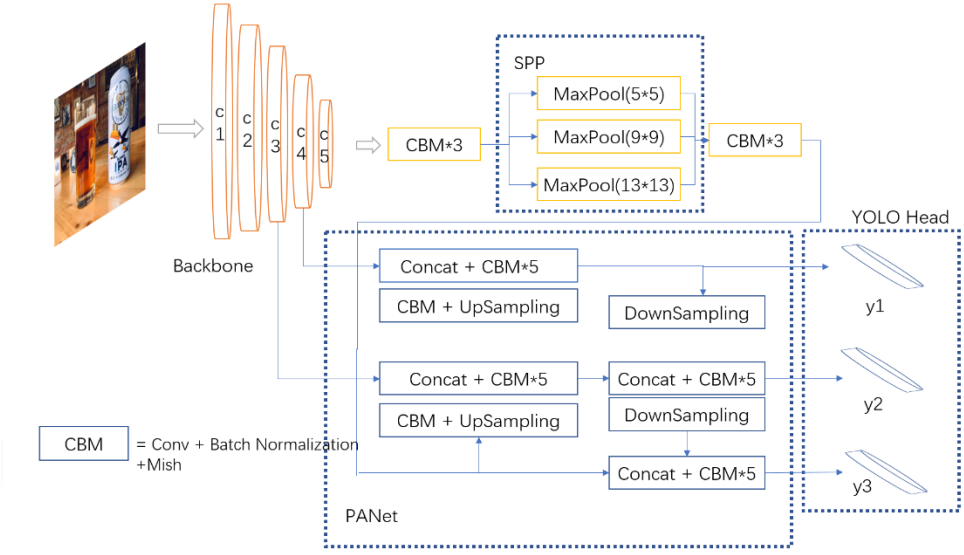

Figure 4. Structure of YOLOv4 model

The description of the structure in Figure 4 proceeds as follows. C1-C5 represents CSPDarknet53, and CBM is a layer composed of Conv, Batch Normalization, and Mish. SPP exports MaxPooling in three types of 5*5, 9*9, and 13*13 after Conv operation, and adds Conv value from the 3 MaxPooling values in Concat and the existing input value then goes through Conv before exporting. Concat plays the role of merging input layers, and Upsampling is a basic library function of pytorch that doubles the number of each array of feature maps in structure values [13]. Downsampling is a basic library function of pytorch that reduces the number of each array in the feature map by two in the structure value. y1-y3 move three different layers and detect and output the same image by changing the size.

3.2.2 SSD model

The SSD model enhances the feature map by adding four convolutional layers to the base network based on VGG-16. SSDs extract feature maps of various scales" to detect objects. We can detect small objects using the large feature maps in the front, and we can detect large objects using the small feature maps in the back. The structure of the SSD model is shown in Figure 5 below.

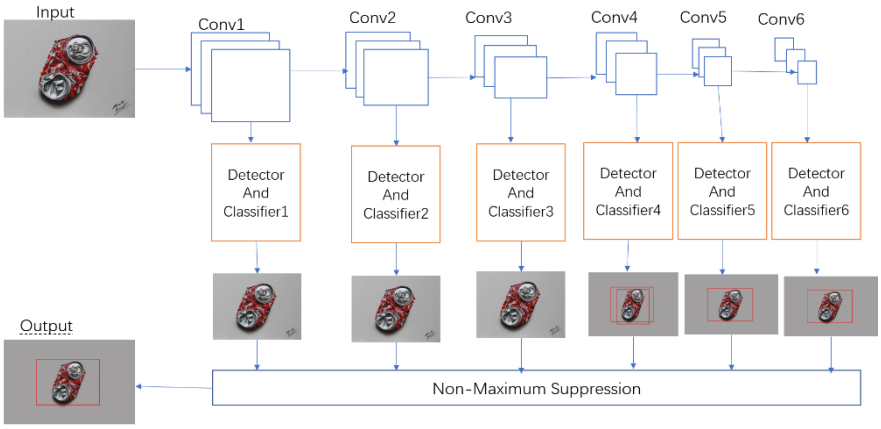

Figure 5. SSD model structure

In order for an image or video to be input to the SSD network, it must be resized to 300*300*3 before sending to the SSD target detection network. When the resized picture is input to the Conv_1 layer of the VGG network, feature map_1 is obtained. Using the result of performing a 3 * 3 convolution kernel through the feature layer obtained from the previous convolutional layer, we can perform maximum pooling and then obtain the feature layer of the layer. By performing a 3 * 3, 1 * 1 convolution kernel on each feature map, it is possible to determine whether there is an object in the proposal box and obtain a regressed result. Enter six candidate boxes in NFS (Non Maximum Suppression) and output the best candidate boxes for object detection.

3.2.3 Faster R-CNN

This paper selected the YOLOv4 algorithm in one stage for comparative analysis between the proposed models and choose the Faster R-CNN algorithm in two stages. Instead of using the sliding window and image pyramid methods like the conventional two-stage method, Faster-RCNN, which represents the two-stage detection model, integrates the structure of the CNN into the overall process of the target detection task. Alternatively, select the Select Search (SS) [14] method to generate a detection frame.

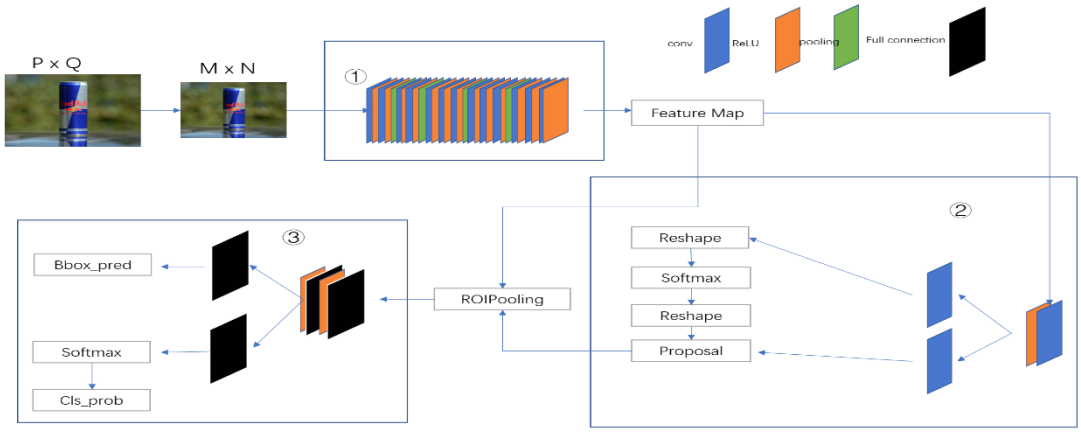

Figure 6. Structure of Faster R-CNN Model

Faster-RCNN uses Region Proposal Networks (RPN) to generate detection frames on the feature maps obtained from the Backbone Network, greatly improving the generation speed. The Faster R-CNN model is shown in Figure 6 below. Figure 6 shows the structure of the Faster RCNN model proposed in this paper. In ①, the Fast R-CNN object detection network is generally a pre-trained CNN feature extraction network similar

to the model of Fast RCNN, and consists of two sub-networks that can be later learned. And the RPN of ② is used to generate the object proposal, and finally ③ is used to predict the actual class of the object. In addition, the differentiating factor of Faster R-CNN is the RPN inserted after the last convolutional layer, which is trained to directly generate local proposals without external mechanisms such as selective search. We then use ROI pooling and similar upstream classifier and bounding box regression with Fast R-CNN.

4. RESULT

Training and testing were carried out using 'wandb', an API that records loss and accuracy when learning a neural network. In this paper, in order to check the loss and accuracy, the training of three models of YOLOv4, SSD, and Faster R-CNN was divided into 1 step, 20 step, 50 step, 100 step, and 300 step, comparative evaluation.

Table 2. Loss values of three models

Division	Loss val of Faster R-CNN		SSD loss val		YOLOv4 loss val	
	Train	Value loss	Train	Value loss	Train	Val loss
1 step	1.0639	0.5646	5.3545	2.3298	86.7438	9.9385
20 step	0.2700	0.3504	1.4030	0.9568	0.5921	0.4184
50 step	0.2081	0.3421	1.1469	0.7676	0.3512	0.2970
100 step	0.1576	0.3022	0.5786	0.6662	0.2460	0.0989
300 step	0.1522	0.2992	0.5810	0.6471	0.2241	0.1220

In this paper, Faster R-CNN, SSD, and YOLOv4 were used to find a suitable model for detecting waste, and each training was repeated 300 times. The loss rate of the progressed learning is the result for the train loss as shown in Table 2.

In the initial stage of learning, step 1, the loss value was 1.064 for Faster R-CNN, 5.355 for SSD, and 86.744 for YoloV4. At step 20, Faster R-CNN was 0.270 and YOLOv4 was 0.592, and SSD slowly dropped to 1.403. In Step50, the loss values of the three models were 0.208 for Faster R-CNN, 1.147 for SSD, and 0.351 for YOLOv4. In Step 100, all three models were stable at about 0.158, 0.579, and 0.246, and in the last Step 300, Faster R-CNN was trained at 0.152, SSD at 0.581, and YOLOv4 at 0.224.

Looking at the training data above, Faster R-CNN can be fitted quickly in about Step 50, and the robustness of the Faster R-CNN model was demonstrated. And from the training data, it can be seen that YOLOv4 has the advantage of fast training speed in model training, but lacks model stability compared to Faster R-CNN. SSD has the slowest training speed, but is more stable compared to the YOLOv4 model.

This paper tested to find an appropriate model for the identification and classification of municipal waste, and considers the detection speed and accuracy to select a network model suitable for the classification of municipal waste. The proposed three models are tested 300 times each in the same data set and environment, and the training results are recorded in 'wandb'.

Figure 7 shows the training and testing trends of the three models Faster R-CNN, SSD and YOLOv4. The training loss value of Faster R-CNN first step is 1.0639, step 20 is 0.2700, step 50 is 0.2081, step 100 is 0.1576, step 300 is 0.1522. The test loss value of the first step is 0.5646, and step 20 is 0.3504, the 50th step is 0.3421, the 100th step is 0.3022, and the 300th step is 0.2992. The training loss value for the first step of SSD is 5.3545, the 20th step is 1.4030, the 50th step is 0.1469, the 100th step is 0.5786, and the 300th step is 0.5810. The first test loss value is 2.3298, the 20th step is 0.9568, and the first step is 0.9568. The 50th step is 0.7676, the 100th step is 0.6662, and the 300th step is 0.6471. The training loss value of the first step of YOLOv4 is 86.7438, the 20th step is 0.5921, the 50th step is 0.3512, the 100th step is 0.2460, and the 300th step is 0.2241. The first step test loss value is 9.9385, the 20th step is 0.4184, The 50th step is 0.2970, the 100th step is 0.0989, and the 300th step is 0.1220.

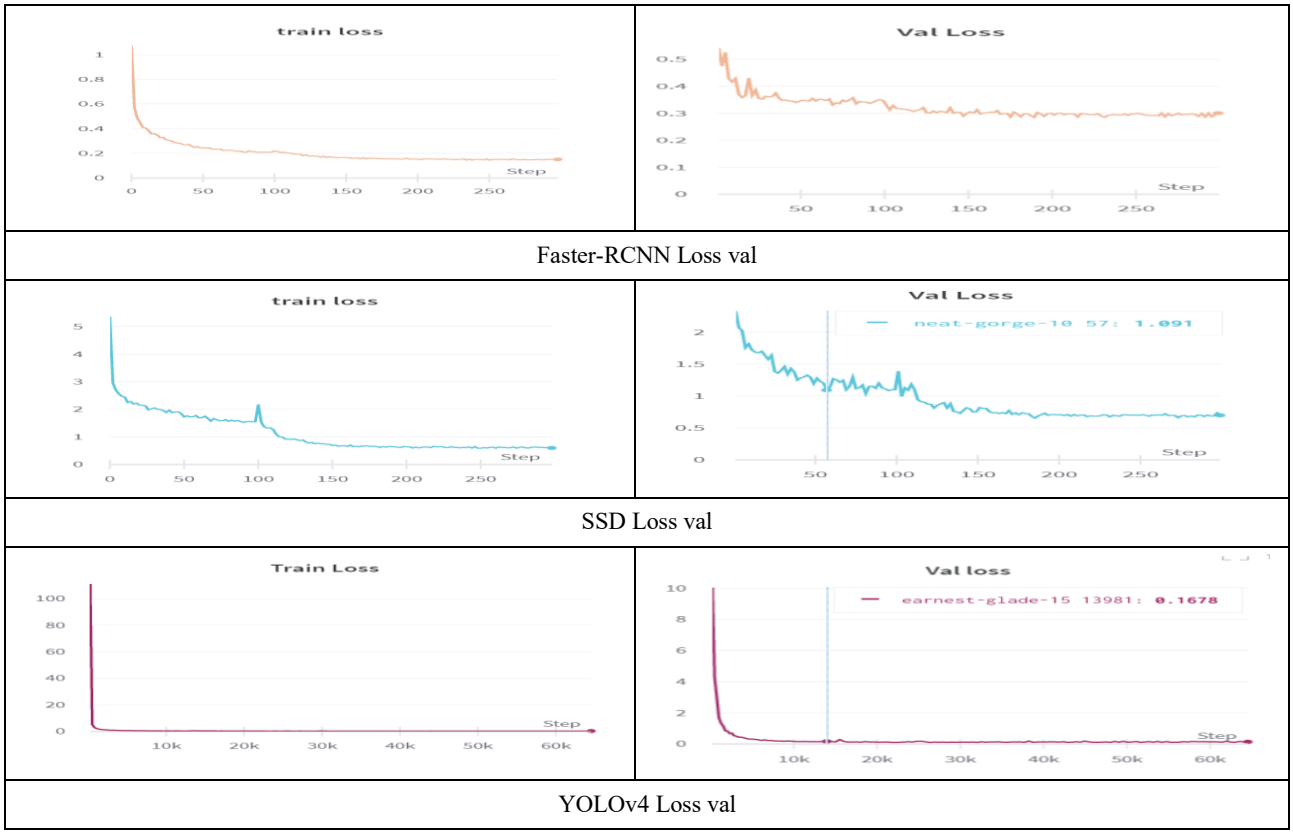

Figure 7. Graph of Learning Results

In order to compare and analyze the Average Precision, F1, Recall, Precision, and mAP of the three models of Faster RCNN, SSD, and YOLO v4 in garbage classification to determine which model is more suitable for garbage classification. The results are shown in Table 3.

Table 3. Results of three models

Division	Indicators	cans	plastics
Results of the YOLOv4 model	Average Precision	94.82%	77.13%
	F1	93.00%	79.00%
	Recall	88.39%	70.63%
	Precision	97.93%	90.36%
	Fps	32-33	
	mAP	85.97%	
Results of the SSD model	Average Precision	99.99%	99.56%
	F1	99.00%	98.00%
	Recall	100.00%	98.81%
	Precision	98.52%	97.65%
	Fps	33-35	
	mAP	99.78%	
Results of the Faster R-CNN model	Average Precision	99.40%	98.83%
	F1	97.00%	96.00%
	Recall	100.00%	99.20%
	Precision	94.35%	99.22%
	Fps	17-20	
	mAP	99.11%	

Table 3 shows the results of the three models proposed in this paper, and the contents are as follows. As a result of YOLOv4, the AP value of can was 94.82%, the value of F1 was 93.00%, Recall 88.39%, Precision 97.93%, and the AP value of plastic was 77.13%, F1 value 79.00%, Recall 70.63%, Precision 90.36 In terms of %, the overall mAP value was 85.97%, which was the lowest when compared with the three models. The SSD result showed that the AP value of can was 99.99%, the value of F1 was 99.00%, Recall 100.00%, Precision 98.52%, and the AP value of plastic was 99.56%, F1 value was 98.00%, Recall 98.81%, Precision 97.65 %, the overall mAP value was 99.78%, which was the highest among the three models. Faster R-CNN results showed that the AP value of can was 99.40%, the value of F1 was 97.00%, Recall 100.00%, and Precision 94.35%, Precision 99.22%, the overall mAP value was 99.11%.

In this paper, we set up three models of Faster-RCNN, SSD, and YOLOv4 to compare and analyze the performance of image detection. As shown in Table 2 and Picture 7, among the three models set in this article, YOLOv4 converges faster, followed by Faster R-CNN, and SSD is the most unstable. In the garbage classification and recognition of can, the neural network models of the three are all adapted to the processing of this data set. Among them, the processing speed of SSD and YOLOv4 is relatively fast and can reach about 30fps in line with industry standards. Although Faster R-CNN has good accuracy, it still lacks in processing speed. In terms of plastics garbage classification and identification, although yolo has a faster processing speed, it is lacking in accuracy compared with SSD and Faster R-CNN models. The other two ensure high accuracy while the processing speed of SSD is much higher than Faster R-CNN. In summary, in the experiment of this paper, SSD is more suitable for processing garbage classification data sets.

5. CONCLUSION

Currently, due to COVID-19, household waste has a lot of impact on the environment due to packaging of food delivery. In this paper, we design and implement Faster-RCNN, SSD, and YoloV4 models for municipal waste detection and classification. The data set searched for two types of plastic type and aluminum can type, which account for a large proportion of household waste, and loss value and accuracy value for household waste classification detection using three models, Faster-RCNN, SDD, and YoloV4. Was comparatively evaluated.

During the study, the YOLOv4 model was selected in one stage, and the Faster R-CNN model was selected in two stages. Faster-RCNN, which represents the two-stage detection model, integrates the entire process of target detection with a structure of CNN different from the existing two-stage method, and selects the Select Search method to generate a detection frame. Faster-RCNN uses Region Proposal Networks (RPN) to generate detection frames on the feature map acquired from the Backbone Network, greatly improving the generation speed.

According this paper, the SSD model cans achieve the best results with an average precision of 99.99%, an average precision of plastics of 97.65%, and a mAP value of 99.78%.

REFERENCES

[1] “Korea resource recirculation Information system”, http://210.104.107.10/rrs/main.do
[2] Abhishek Gupta, Alagan Anpalagan, Ling Guan, Ahmed Shaharyar Khwaja, “Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues,”ELSEVIER, scienceDirect, Vol. 10, July 2021. DOI: https://doi.org/10.1016/j.array.2021.100057.
[3] X.L. Xie, Recognition and Positioning of Sorting Robot Based on Deep Learning, An-Hui Engineering University, CN, 2019.
[4] Y.F.Zhang. Research and Implementation of Person Detection System Based on Deep Learning, Beijing University of Posts and Telecommunications, CN, 2020.
[5] https://blog.csdn.net/Payforr/article/details/107392539.
[6] Simonyan, K. and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” CoRR, pp.1409-1556, 2015.
[7] H.Xu, D.G.Yang, Q.Q.Jiang, L.J.Lin, “Improvement of Lightweight Vehicle Detection Network Based

on SSD" Computer Engineering and Applications, pp.1-10, June 2021.
[8] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017. DOI: 10.1109/TPAMI.2016.2577031.
[9] "ZhiHu", ZhiHu, last modified, accessed Sep 06, 2021, https://zhuanlan.zhihu.com/p/31426458.
[10] "dobong-gu Seoul", Korea Cleaning Administration Website, last modified June 02, 2021, accessed Sep 06, 2021, https://www.dobong.go.kr/subsite/waste/Contents.asp?code=10007358
[11] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6022-6031, 2019. DOI: 10.1109/ICCV.2019.00612.
[12] HE K, ZHANG X, REN S, et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, No. 9, pp.1904-1916, 2015. DOI: 10.1109/TPAMI.2015.2389824
[13] "PyTorch", PyTorch, last modified, accessed June 02, 2021, https://pytorch.org/docs/master/generated/torch.nn.Upsample.html
[14] Uijlings, J. R. R, VanDesande, K. E. A, Smeulders, A. W. M. and Gevers, T. "Selective Search for Object Recognition," International Journal of Computer Vision, Vol. 104, No. 2, pp. 154-171, 2013. DOI: 10.1007/s11263-013-0620-5