An Improved Ensemble Learning-Based Approach for Retail Product Recognition

Huei-Yung Lin , Po-Yu Hsieh , Sen-Yih Chou , Chin-Chen Chang *

Posted Date: 17 November 2023

doi: 10.20944/preprints202311.1141.v1

Keywords: ensemble learning; retail product recognition; convolutional neural network; unmanned store

Article

An Improved Ensemble Learning-Based Approach for Retail Product Recognition

Huei-Yung Lin [1], Po-Yu Hsieh [2], Sen-Yih Chou [3] and Chin-Chen Chang [4,*]

[1] Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; lin@ntut.edu.tw
[2] Department of Electrical Engineering, National Chung Cheng University, Chiayi 621301, Taiwan; hs0920920832@gmail.com
[3] Center for Measurement Standards, Industrial Technology Research Institute, Hsinchu 310401, Taiwan; yih@hotmail.com
[4] Department of Computer Science and Information Engineering, National United University, Miaoli 360302, Taiwan; ccchang@nuu.edu.tw
* Correspondence: ccchang@nuu.edu.tw

Abstract: Due to the recent trend of unmanned economy, retail stores have gradually reduced their service and cashier manpower. The retail product recognition becomes one of key issues for unmanned shopping. Although the success of deep neural networks has made object recognition feasible in a variety of applications, it still struggles to perform well on a large number of classes of retail products. In this paper, we propose an improved ensemble learning-based approach for retail product recognition. In the proposed approach, the object classification network is first improved through feature extraction and block attention. An ensemble model is then built by integrating multiple network models and using loss selection as model weights. In the experiments, the feasibility of our ensemble learning-based approach has been evaluated through many production items. The results demonstrate the effectiveness of the proposed approach compared with previous retail product recognition methods.

Keywords: ensemble learning; retail product recognition; convolutional neural network; unmanned store

1. Introduction

As declining birth rates and high labor costs become a growing problem around the world, traditional retailers are forced to reduce the manpower and cost requirements of store management. Some common methods include using internet of things (IoT) monitoring systems to ensure product health and safety [1] and deploying surveillance cameras for security purposes [2]. However, most current methods do not completely eliminate the need for on-site manpower. Therefore, the concept of unmanned stores in the retail industry has often been proposed and studied [3].

An unmanned store generally refers to a shopping model where there are no service personnel in the store. Customers can view the products and complete the purchase by themselves. To realize unmanned shopping in unmanned stores, it needs to rely on many technologies of artificial intelligence, including e-commerce detection, recognition, and big data analysis [4]. The most famous unmanned retail store is Amazon Go, launched in 2018. It uses technologies such as sensing, vision, machine learning, and cloud computing to achieve automated transactions [5]. Customers can purchase through the Amazon Go app and the store does not require any clerks. However, there are still many challenges in terms of customer experience, especially the ease of use of viewing product items. Current methods of scanning barcodes are often considered unfriendly and sometimes difficult for flexible bag packaging. Therefore, an important task is to easily recognize product items.

Due to the recent success of using deep neural networks for object recognition, it is possible to detect and recognize product items using image-based methods. Convolutional neural networks (CNN) have been widely used for object classification. Existing literature reports some promising

 doi:10.20944/preprints202311.1141.v1

results for different application scenarios [6-7]. A major problem with these techniques for recognizing retail products is the excessive number of classes. There can be hundreds of different retail products in the same class on the racks which makes recognition a challenging task. Furthermore, similar designs and patterns used for specific types of retail products further complicate feature extraction and network training [8].

In order to solve the problem of significant performance degradation when the number of classes increases, it is necessary to establish a more complex classification network architecture. In addition, a sufficient number of samples are needed to extract representative features for network training. However, neither is a trivial task and usually involves improving upon existing methods. An important strategy is to use the concept of ensemble learning to improve the recognition rate by combining different classification algorithms [9,10]. In most supervised learning techniques, data classification is achieved by training a classifier using ground-truth annotations. However, each learning method may have different classification effects on different datasets. Therefore, through the concept of ensemble learning, multiple learning algorithms can be used to obtain better prediction results than a single learning algorithm.

Due to high labor costs in many places of the world, retailers have to reduce the manpower and costs of store management. This includes basic retail product recognition for checking product items and transactions. However, current object recognition technologies cannot accurately identify a variety of retail products. Image recognition of a large number of different samples is an emerging problem. It is different from traditional object recognition, which only considers a smaller number of classes and limited sample diversity. Generally, in classification algorithms, objects with similar attributes or functions belong to the same class. However, recognizing the same object using images captured from different lighting and viewpoints requires recognizing the unique object. In this regard, the recognition of many retail products is similar to face recognition, but more challenging.

In this paper, we propose an ensemble learning-based approach for retail product recognition. We first improve the object classification network through based on feature extraction and block attention. We then build an integrated model by integrating multiple network models and using loss selection as model weights. Our approach demonstrates the feasibility of recognizing up to fifty retail products.

2. Related Works

According to the combination strategy of ensemble learning, current methods can be divided into three categories. The first method is the boosting algorithm, which combines multiple weak classifiers into a strong classifier. It requires the learning error rate of the weak classifier to update the weights of the training samples. The purpose is to increase the weight of misclassified data by the old classifiers and then train the new classifiers. This will provide the new classifiers with the features to learn misclassified samples, improve the classification effect, and obtain an ensemble model with high recognition accuracy. The second method is the bagging algorithm. It randomly selects n samples from the training data each time and puts them back. Due to random sampling, subsets of samples collected multiple times are often different from each other and from the original training data set. Repeating this process m times can produce m weak classifiers. Then, average, voting or other methods are used for ensemble learning. The third and most widely used neural network ensemble learning technique is stacking. In this method, the original data set is used to train a first weak classifier for prediction, and then the labeled data set of the second weak classifier is used to produce an ensemble model [11].

Li et al. [12] proposed a method that not only considers the quality of the ensemble model based on the overall mean average precision (mAP), but also selects different CNN models based on the class average precision (AP). The advantages of the models are complementary to each other and AP is used as a weight to determine confidence. Then, it is followed by the voting scheme to improve the performance of the ensemble model. Alam et al. [13] presented a dynamic ensemble learning algorithm to design and train the ensemble of neural networks. Their approach can automatically create ensemble architectures that maintain the accuracy and diversity of the neural networks, as well

 doi:10.20944/preprints202311.1141.v1

as the minimum number of parameters specified by the designer. Their method can achieve good generalization ability. Zhu et al. [14] proposed a method to solve the imbalance problem of training data through geometric structure ensemble learning method. First, their method generated a hypersphere through Euclidean metric to divide and eliminate redundant majority samples. Second, this method learned a basic classifier to enclose a small number of samples. Therefore, their approach can achieve higher efficiency during the training process. Finally, the remaining samples were used to train the next sample until the entire training process is completed.

Santra et al. [15] proposed an end-to-end annotation-free machine vision system for detecting products on the rack. Their system consisted of three modules: exemplar-driven region proposal, classification, and non-maximal suppression of region proposals. First, they estimated the scale of the rack images relative to product template images. Their system then used the estimated scale to generate potential object regions. Finally, a convolutional neural network (CNN) was used to classify potential object regions. George et al. [16] presented an approach for per-sample multi-label image classification of products in retail store images. Their method used discriminative random forests, deformable dense pixel matching, and genetic algorithm optimization to achieve high efficiency. Furthermore, they performed product image search using tagging tools for multi-label retail product image classification. Their approach achieves good results in both accuracy and efficiency. Wang et al. [17] proposed a fine-grained classification algorithm for retail product recognition based on self-attention destruction and construction learning. Their method uses self-attention techniques to destroy and construct image information in an end-to-end manner. Therefore, we can compute accurate fine-grained classification predictions and large informative regions during inference. The results show that their approach performs better than other fine-grained classification methods.

Osokin et al. [18] proposed a one-shot target detection algorithm that jointly performs localization and recognition. First, they determined correspondences using dense correlation matching of learned local features. Then, a feedforward geometric transformation model was introduced to align the features. Finally, they used bilinear resampling of the correlation tensor to compute detection scores for aligned features. In particular, their method can detect unseen classes and outperforms multiple baselines. Geng et al. [19] presented a fine-grained product classification approach using feature-based matching and one-shot deep learning. They first detected candidate regions of product instances and used repeated features in product images as coarse class labels for these regions. Then, they generated attention maps to guide the classifier to effectively improve the accuracy of fine-grained grocery product recognition. Their approach had good adaptability and can be used to improve existing classifiers. Tonioni et al. [20] proposed an algorithm for recognizing grocery products on store shelves. First, they used a class-agnostic object detection technique to extract region proposals containing individual product items that appear on shelf images. Second, they used *K*-nearest neighbors (*K*-NN) to perform product recognition separately for each proposed region. Finally, they performed a final refinement of the recognition output by re-ranking the top *K* proposals provided by the similarity search.

3. The Proposed Approach

In our approach, we first improve the RefineDet network [21] by modifying feature extraction and attention modules. We then add the attention modules to the RefineDet network model for optimization. Finally, we build an ensemble learning model to achieve better performance.

3.1. Improved RefineDet

In the original network model of RefineDet, VGG16 is used for feature extraction [22] and the single-shot detector (SSD) is used as the backbone architecture [23]. It combines region proposal network (RPN) and feature pyramid network (FPN) to ensure recognition rate under single small target detection [24]. The RefineDet network model consists of three main components, the anchor refinement module (ARM), the transport connection block (TCB), and the object detection module (ODM). We make modifications on both feature extraction and attention modules of RefineDet to optimize overall performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

First, feature extraction module is an important part for RefineDet to recognize images. Three different networks, ResNeSt [25], RegNet [26], and RexNet [27], are integrated into the feature extraction module of RefineDet. The improved network model is shown in Figure 1, which still retains the original idea of using ARM, TCB and ODM modules. The orange shaded areas change to ResNeSt, RegNet, and RexNet. Next, the attention modules are incorporated into RefineDet to improve feature extraction of convolutional layers.

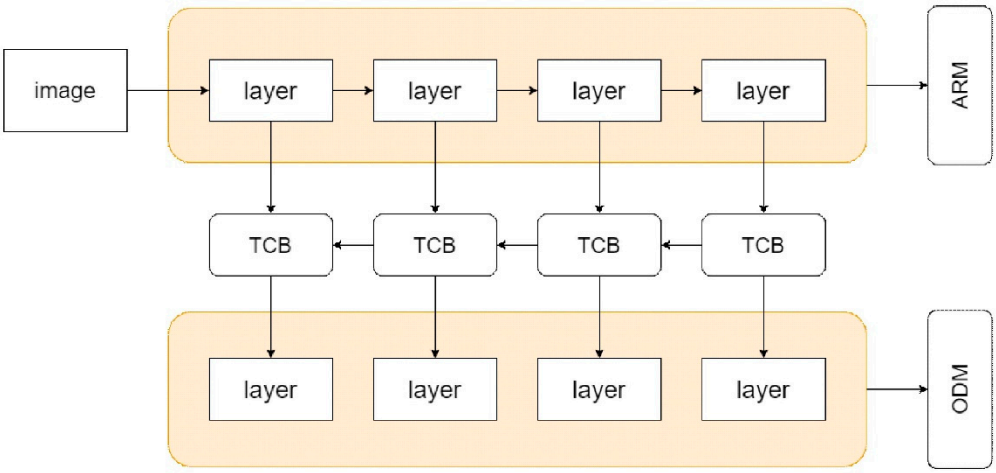
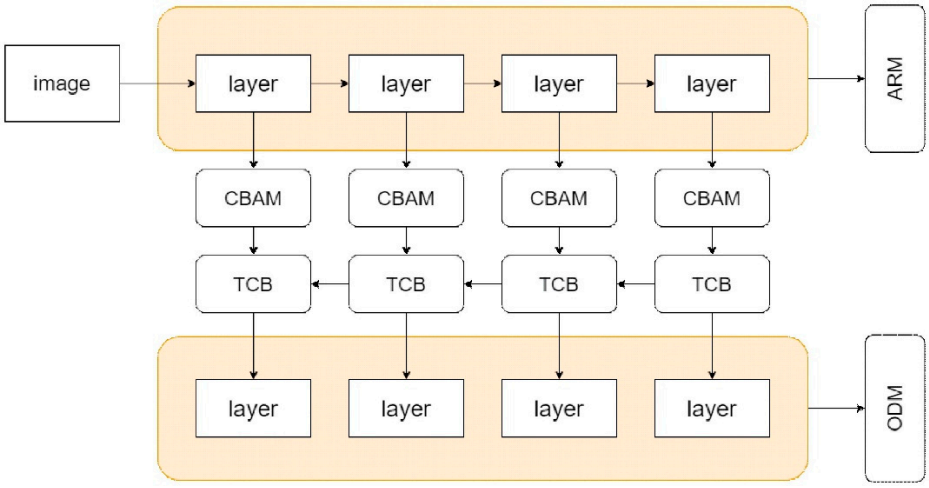


Figure 1. The improved network model.

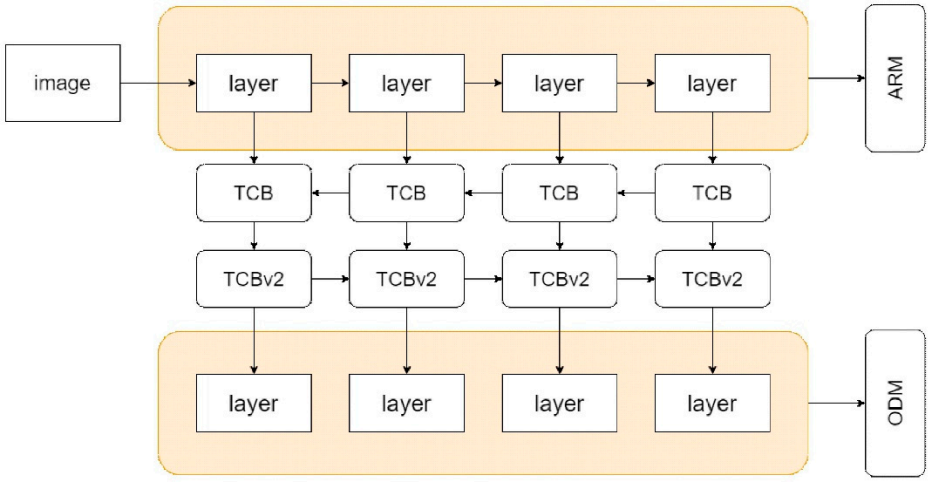
3.2. Attention Module

As shown in Figure 2, three attention modules are used: convolutional block attention module (CBAM), TCBv2, and enhanced block (EMB). The CBAM is mainly divided into channel and spatial attention modules [28]. The model uses max pooling to obtain feature differences in images and average pooling to extract common features for global feature learning. TCBv2 is based on BATCB [29], which joins the ideas of PANet [30] and includes FPN with bottom-up path. This is then integrated into the TCB module in RefineDet. EMB [31] contains attention streams and feature maps. It provides a feature-enhancing approach that improves detection capabilities without reducing processing speed.

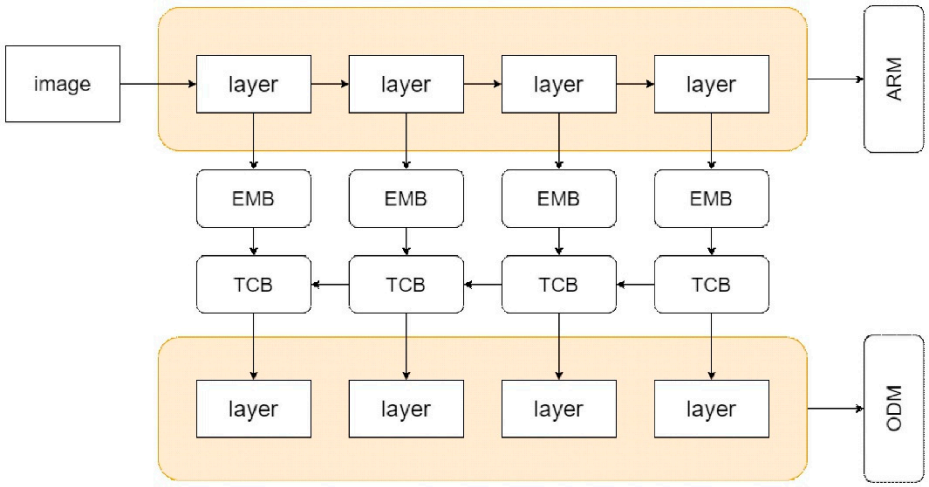


(a)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1



(b)



(c)

Figure 2. Three attention modules, CBAM, TCBv2, and EMB, are adopted to optimize RefineDet. (a) optimized RefineDet model with CBAM, (b) optimized RefineDet model with TCBv2, (c) optimized RefineDet model with EMB.

3.3. Ensemble Learning Model

In order to build our ensemble learning model and improve the recognition rate, we replaced VGG16 with ResNeSt, RegNet and RexNet in RefineDet for feature extraction. The three models are trained separately, and their respective mAP is used as the basic recognition rate for integration and improvement. Our ensemble learning approach is based on the training losses of individual models. The feature extraction layers of different classification networks use the same training data. This process is performed in parallel, and the weighted sum of the training losses is used for performance evaluation of the ensemble model. We first set the weights based on the mAP of the base classification model. Although this configuration can provide good recognition rates, if the network training is updated with changed training data, the weights need to be adjusted accordingly.

To solve this problem, an automatic weight adjustment technique for model ensemble is proposed. It consists of loss selection and joint loss calculation. As shown in Figure 3, the loss of each model can be calculated by selecting one from n batches of training at each epoch. The obtained model loss is used to adjust the contribution ratio among the underlying network structures. In general, the higher loss corresponds to the poorer recognition result and a lower model weight is assigned

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

accordingly. For the joint loss computation, the ensemble loss is calculated by the weighted sum of model training losses.

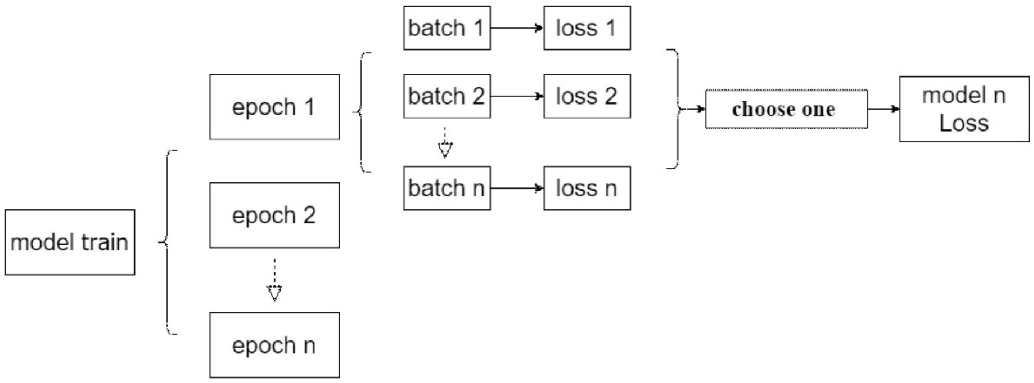


Figure 3. The loss for each model can be derived by selecting one from n batches of training at each epoch.

As shown in Figure 4, the weight ratio of each epoch during model training is derived based on the reciprocal of the model loss in the previous epoch. For a more general formulation with n base network models, the ensemble loss is given by

$$L^i = \sum_{j=1}^{n} W_j^i \cdot L_j^i \tag{1}$$

where

$$W_j^i = \frac{1}{L_j^{i=1} \sum_{j=1}^{n} \left(1/L_j^{i-1}\right)} \tag{2}$$

and L_j^i is the training loss of the network model j at the i-th epoch. For the initialization of the first training epoch, assign the same weights to calculate the ensemble loss. Finally, the ensemble loss is used for backpropagation at each epoch.

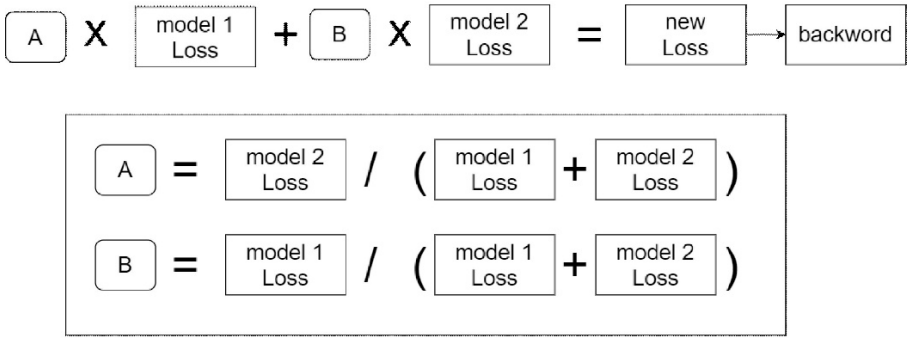


Figure 4. Schematic diagram of ensemble loss calculation. During model training, the weight ratio for each epoch is derived based on the reciprocals of the model losses in the previous epoch.

4. Results

We evaluated the proposed ensemble learning-based approach and network model for retail product recognition with real scene images. Product images were captured by a camera (GoPro HERO 7) with a resolution of 1920×1440, mounted on the top of the fixed bracket and viewing downward. Many products, including various snacks, were placed in different orientations and lighting for image acquisition. The calculations for retail product recognition run on a PC with an Intel Core i5 8600 CPU and an NVIDIA Geforce GTX1080 GPU.

Figure 5 showed several sample images acquired in our experiments and used as the training data. There were 50 product items in the training dataset and each product item had 200 images. As shown in Figure 6, the captured test images were slightly different from the training samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

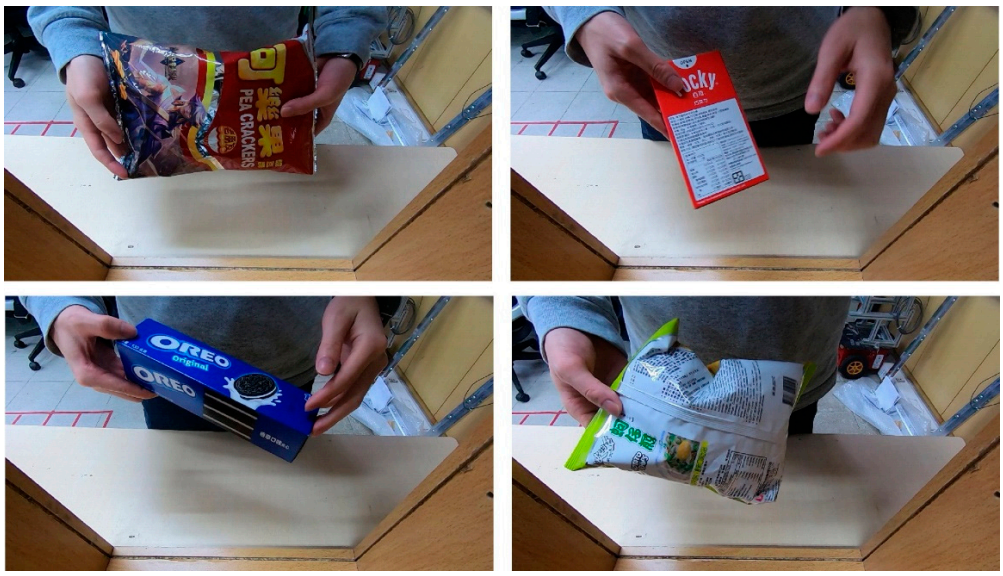


Figure 5. Several images acquired in our experiments and used as the training data.

We replaced the feature extraction layer of RefineDet with ResNeSt50, RegNet and RexNet. In initial tests, a total of 500 images of 10 retail products were used for evaluation. Table 1 showed the mAP results of comparing the original RefineDet and modified RefineDet network models. The mAP increased by approximately 30% for all three conditions. Figure 7 showed the results using the modified RefineDet network models of three feature extraction layers ResNeSt50, RegNet, and RexNet. Although some snacks were blocked by hands, the three modified RefineDet network models can still recognize the product classes and locations. Compared with the original RefineDet network model using VGG16, the recognition results of the three modified RefineDet network models were more stable.

Table 2 showed the results obtained using the proposed ensemble learning-based approach. There were three ensemble models for comparison, namely RexNet/ResNeSt50, ResNeSt50/RegNet, and RexNet/RegNet, trained using the same dataset with loss-based weight adjustment. The mAP results were improved to about 90% with our ensemble models.

Figure 6. The testing images were captured slightly different from the training samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

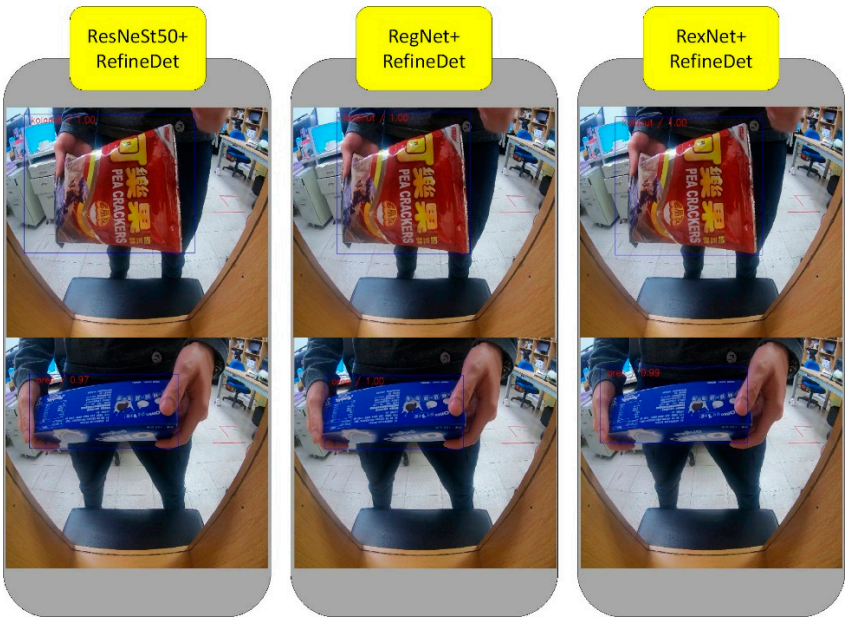


Figure 7. The results using the modified RefineDet network models of three feature extraction layers ResNeSt50, RegNet, and RexNet.

We then gradually included more product items for model training and testing up to 50 classes with a total of 10,000 images. Table 3 showed the test results using 500 images for each class. The average recognition rate drops significantly.

Table 1. The mAP comparison for different modified RefineDet network models.

Model	RefineDet	w/ResNeSt50	w/RegNet	w/RexNet
pocky	86%	79%	85%	74%
cocopuff	85%	81%	80%	92%
doritos	73%	75%	58%	79%
noodles	64%	92%	91%	94%
greengood	51%	87%	96%	88%
kolanut	43%	84%	77%	84%
oysterchip	35%	81%	82%	86%
oricracker	34%	77%	78%	86%
oreo	18%	77%	88%	91%
cookie	18%	71%	64%	94%
mAP	50.84%	80.44%	79.97%	86.83%

Table 2. The results obtained using the proposed ensemble learning-based approach.

Model	Res50+Rex loss	Res50+Reg loss	Rex+Reg loss
pocky	78%	88%	83%
cocopuff	92%	94%	92%
doritos	85%	69%	91%
noodles	96%	96%	88%
greengood	95%	98%	86%
kolanut	87%	91%	82%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

oysterchip	96%	96%	82%
oricracker	87%	94%	88%
oreo	90%	98%	100%
cookie	98%	100%	100%
mAP	90.42%	92.37%	89.23%

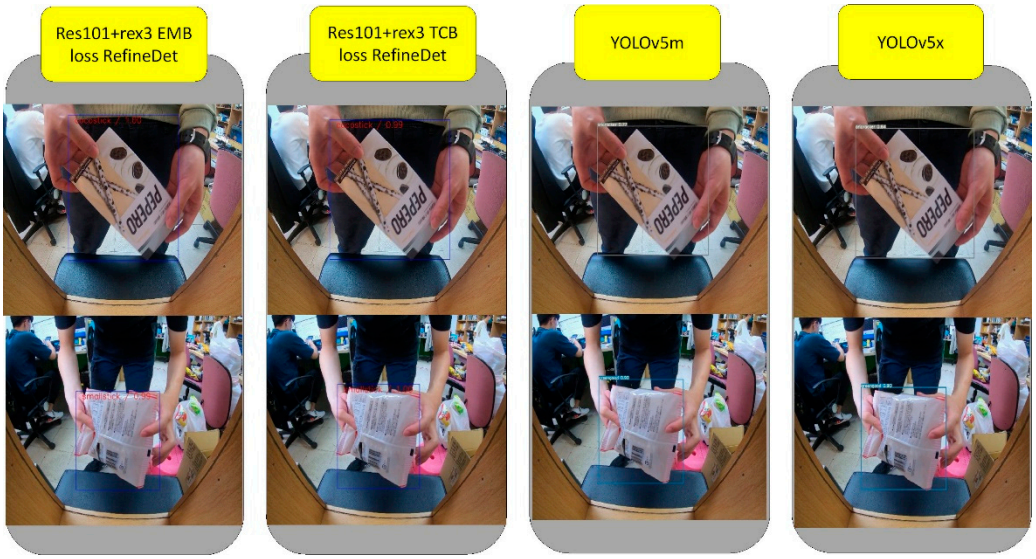
Table 3. The testing results using 500 images for each class for improved RefineDet network models.

Model	Res50+Rex loss	Res50+Reg loss	Rex+Reg loss
mAP	52.1%	51.08%	47.41%

In order to further improve the recognition rate of various product items, we used the ensemble model with the best mAP (i.e. Res50/Rex loss) for improvements. Retrain the network with an additional 100 images containing classes with mAP less than 60%. In the modification of the basic network model, we replaced the initial feature extraction module of ResNeSt50 with ResNeSt101. RexNet was changed to the width of 3 (Rex3). Furthermore, the number of convolutional layers and kernel size were increased. Res101/Rex3 CBAM was constructed by adding an attention module. It used pooling to enhance feature extraction. Res101/Rex3 TCB was modified, replacing the TCB in RefineDet with BATCB, enhancing the data conversion of ARM and ODM, and fusing the features from shadow and deep layers. Res101/Rex3 EMB was built by adding EMB modules to optimize ARM-derived features. This ensemble model provided the best results in the experiments.

Table 4 showed the comparison of different ensemble models and YOLOV5 for 50 classes of retail products. As shown in Table 4, the mAP of the ensemble model Res101/Rex3 was quite stable for each class. Both Res101/Rex3 TCB and Res101/Rex3 EMB improved mAP by more than 5% compared to the original Res101/Rex3 model. The results showed that our ensemble model provided better mAP and can recognize challenging cases such as “smallstick” and “cocostick” (which have very low mAPs from YOLOv5). As shown in Figure 8, the retail products on top and bottom were “cocostick” and “smallstick”, respectively. For YOLOv5m and YOLOv5x, they were recognized as “greengood” and “oricracker”, respectively.

Table 5 listed frames per second (FPS), the number of parameters, and the training time for comparing the proposed ensemble models and YOLOv5. The results indicated that the performance of the ensemble model is not good compared to YOLOv5. This is because the idea of ensemble learning is to integrating multiple networks for improving recognition rate by sacrificing FPS, the number of parameters, and training time. It illustrated the trade-off between accuracy and resources of the ensemble models.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

Figure 8. The comparison of Res101+rex3 EMB loss RefineDet, Res101+rex3 TCB loss RefineDet, YOLOv5m, and YOLOv5x. The retail products on top and bottom were "cocostick" and "smallstick", respectively.

Table 4. The comparison of different ensemble models and YOLOV5 for 50 classes of retail products.

Network model	Res101+Rex3 loss RefineDet	Res101+Rex3 CBAM loss RefineDet	Res101+Rex3 TCB loss RefineDet	Res101+Rex3 EMB loss RefineDet	YOLOv5m	YOLOv5x
kido	86%	98%	88%	88%	46%	46%
cookie	92%	93%	94%	62%	92%	92%
applevinegar	90%	92%	90%	90%	56%	8%
cocostick	86%	96%	78%	98%	16%	8%
pestochip	85%	81%	86%	81%	39%	16%
sweetpotato	76%	76%	82%	70%	74%	56%
oreo	86%	90%	90%	100%	30%	96%
greengood	72%	84%	78%	68%	47%	51%
cheesechip	84%	100%	92%	98%	82%	98%
crispyflute	66%	62%	84%	91%	66%	78%
skewernoodles	84%	57%	93%	64%	96%	98%
nenecookie	76%	88%	92%	92%	76%	84%
coconutcookie	56%	56%	76%	78%	42%	16%
oricracker	82%	66%	84%	78%	84%	80%
ritzcracker	82%	93%	87%	86%	60%	58%
tomatofries	67%	66%	72%	80%	96%	94%
purpleairwaves	86%	81%	74%	75%	57%	72%
spicydoritos	68%	85%	81%	92%	94%	94%
lonelygod	81%	81%	94%	84%	66%	82%
cheetos	62%	71%	66%	70%	89%	86%
goodtimechoc	74%	76%	76%	87%	52%	80%
widechip	82%	86%	86%	77%	62%	51%
biscoff	86%	84%	74%	94%	40%	40%
cocopuff	73%	61%	73%	92%	84%	73%
bananapocky	84%	85%	84%	88%	82%	74%
oysterchip	60%	67%	58%	56%	85%	82%
nutritionalchip	80%	80%	90%	90%	26%	62%
purpleqq	85%	92%	86%	83%	96%	98%
nutchocolate	94%	86%	96%	97%	84%	96%
spicykaramju	84%	72%	90%	78%	58%	58%
milkyokan	64%	61%	64%	52%	42%	52%
seaweedcookie	70%	66%	70%	83%	73%	92%
crispynoodles	80%	72%	75%	52%	40%	39%
seaweedsalt	74%	75%	82%	74%	67%	87%
squidcracker	64%	71%	88%	94%	86%	92%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2023 doi:10.20944/preprints202311.1141.v1

yakultcrispyflute	81%	86%	83%	91%	96%	96%
milkcandy	58%	38%	46%	67%	59%	86%
seafoodcracker	64%	73%	77%	79%	64%	87%
kolanut	47%	35%	69%	64%	88%	89%
pocky	92%	76%	84%	84%	78%	92%
bakecookie	49%	54%	46%	76%	81%	75%
newpie	46%	57%	57%	86%	83%	98%
haribosoft	54%	72%	74%	92%	84%	79%
darkchocolate	59%	51%	62%	77%	82%	65%
blackcookie	80%	55%	69%	76%	62%	78%
cocopeacockroll	47%	44%	53%	86%	60%	88%
smallstick	69%	53%	89%	75%	0%	0%
strawberrycrisp	47%	33%	62%	68%	98%	98%
minicrispdoritos	73%	37%	72%	67%	85%	80%
brownsugarbar	60%	59%	94%	47%	87%	96%
mAP	**72.94%**	**71.49%**	**78.24%**	**79.53%**	**67.96%**	**71.93%**

Table 5. The FPS, the number of parameters, and the training time for the proposed ensemble models and YOLOv5 for comparison.

Model	Res101+Rex3 EMB	Res101+Rex3 TCB	YOLOv5m	YOLOv5x
FPS	2.13	2.02	6.87	6.36
Parameters (M)	176.82	184.4	21.4	87.7
Training time (hr)	60	60	36	36

5. Conclusions

We have proposed an ensemble learning-based approach for retail production recognition. The classification network, RefineDet, is first modified using new feature extraction and attention modules to improve the basic recognition rate. An ensemble models are then built by combining the network models with a weighted loss. The proposed ensemble learning-based technique is evaluated using 50 retail products. Compared with state-of-the-art networks, our approach demonstrates the feasibility of recognizing a large number of object classes.

In future research, we will try to add more ensemble learning methods for analysis and upgrade the basic recognition model to more than three. In addition, we will propose an ensemble model that is better than the current method. Finally, we will add more classes for the proposed approach to test whether the accuracy is maintained.

Author Contributions: Methodology, P.-Y.H, S.-Y.C and H.-Y.L.; Supervision, H.-Y.L. and C.-C.C.; Writing—original draft, P.-Y.H and S.-Y.C; Writing—review & editing, H.-Y.L. and C.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: The support of this work in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2221-E-194-004 and Center for Measurement Standards, Industrial Technology Research Institute is gratefully acknowledged.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. M.H. Ahmadzadegan, M. Mohammadzadeh, G. Eftekharnejad, H. Ghorbani, Intelligent monitoring systems for transportation of perishable products based internet of things (iot) technology. In Proceedings of 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), Gwalior, India, 10–12 April 2020.
2. Q. Xu, W. Zheng, X. Liu, P. Jing. Deep learning technique based surveillance video analysis for the store. Applied Artificial Intelligence, vol. 34, no. 14, pp. 1055–1073, 2020.
3. H.C. Wu, C.H. Ai, C.C. Cheng. Experiential quality, experiential psychological states and experiential outcomes in an unmanned convenience store. Journal of Retailing and Consumer Services, vol. 51, pp. 409–420, November 2019.
4. X. Hu, Y. Yang, L. Chen, S. Zhu. Research on a prediction model of online shopping behavior based on deep forest algorithm. In Proceedings of 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 28–31 May 2020.
5. N. Waranugraha and M. Suryanegara. The development of iot-smart basket: Performance comparison between edge computing and cloud computing system. In Proceedings of 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), Yogyakarta, Indonesia, 15–16 September 2020.
6. W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, vol. 29, no. 9, pp. 2352–2449, 2017.
7. P. Tang, X. Wang, B. Shi, X. Bai, W. Liu, Z. Tu. Deep fishernet for image classification. IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 7, pp. 2244–2250, 2018.
8. V.L. Tran, H.Y. Lin, H.W. Liu. Mbnet: A multi-task deep neural network for semantic segmentation and lumbar vertebra inspection on x-ray images. In Proceedings of the Asian Conference on Computer Vision, virtual Kyoto, November 30–December 4, 2020.
9. H. Liu and S.M. Chen. Multi-level fusion of classifiers through fuzzy ensemble learning. In Proceedings of 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 08–09 December 2018.
10. P.Y. Hsieh, H.Y. Lin, S.Y. Chou. Ensemble learning for retail product recognition with a large number of classes. In Proceedings of 2022 IEEE International Conference on Systems, Man and Cybernetics (SMC 2022), Prague, Czech Republic, 9–12 October 2022.
11. O. Sagi and L. Rokach. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, 2018.
12. J. Lee, S.K. Lee, S.I. Yang. An ensemble method of cnn models for object detection. In Proceedings of 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 17–19 October 2018.
13. K.R. Alam, N. Siddique, H. Adeli. A dynamic ensemble learning algorithm for neural networks. Neural Computing and Applications, vol. 32, no. 12, pp. 8675–8690, 2020.
14. Z. Zhu, Z. Wang, D. Li, Y. Zhu, W. Du. Geometric structural ensemble learning for imbalanced problems. IEEE Transactions on Cybernetics, vol. 50, no. 4, pp. 1617–1629, 2018.
15. B. Santra, A.K. Shaw, D.P. Mukherjee. An end-to-end annotation-free machine vision system for detection of products on the rack. Machine Vision and Applications, 32, article number 56, March 2021.
16. M. George and C. Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. Proceedings of the European Conference on Computer Vision 2014, Zurich, Switzerland, September 6–12, 2014.
17. W. Wang, Y. Cui, G. Li, C. Jiang, S. Deng. A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. Neural Computing and Applications, vol. 32, iss. 18, pp. 14613–14622, 2020.
18. A. Osokin, D. Sumin, V. Lomakin. Os2d: One-stage one-shot object detection by matching anchor features. In Proceedings of the European Conference on Computer Vision 2020, Glasgow, UK, August 23–28, 2020, pp. 635–652.
19. W. Geng, F, Han, J. Lin, L. Zhu, J. Bai, S. Wang, L. He, Q. Xiao, Z. Lai. Fine-grained grocery product recognition by one-shot learning. In Proceedings of the ACM International Conference on Multimedia 2018, Seoul, Korea, 22–26 October 2018, pp. 1706–1714.
20. A. Tonioni, E. Serra, L.D. Stefano. A deep learning pipeline for product recognition on store shelves. In Proceedings of the International Conference on Image Processing, Applications and Systems 2018, Sophia Antipolis, France, 12–14 December 2018.
21. S. Zhang, L. Wen, X. Bian, Z. Lei. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, 18–22 June 2018, pp. 4203–4212.

22. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
23. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision 2016 (ECCV 2016), Amsterdam, The Netherlands, 11–14 October 2016.
24. T.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, Hawaii, USA, 21–26 July 2017.
25. Z. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A. Smola. Resnest: Split-attention networks. In Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, New Orleans, LA, USA, 19–20 June 2022.
26. I. Radosavovic, R.P. Kosaraju, R. Girshick, K. He, P. Dollár. Designing network design spaces. In Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 13–19 June 2020.
27. D. Han, S. Yun, B. Heo, Y. Yoo. Rexnet: Diminishing representational bottleneck on convolutional neural network. arXiv 2020, arXiv:2007.00992.
28. S. Woo, J. Park, J.Y. Lee, I.S. Kweon. Cbam: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
29. H. Xie and Z. Wu. A robust fabric defect detection method based on improved refinedet. Sensors, vol. 20, no. 15, article 4260, 2020.
30. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia. Path aggregation network for instance segmentation. In Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
31. H.T. Choi, H.J. Lee, H. Kang, S. Yu, H.H. Park. Ssd-emb: An improved ssd using enhanced feature map block for object detection. Sensors, vol. 21, no. 8, article 2842, 2021.