2022 4th International Conference on Communications, Information System and Computer Engineering (CISCE) | 978-1-6654-9848-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CISCE55963.2022.9851085

Intelligent Retail Settlement Platform based on Image Retrieval

Xin Yan
School of Information Science and Engineering, East China University of Science and Technology
Shanghai 200237, P.R. China
yx20001210@gmail.com

QingChun Hu
School of Information Science and Engineering, East China University of Science and Technology
Shanghai 200237, P.R. China
* Corresponding author: hqingchun@ecust.edu.cn

XiaoYue Huang
School of Information Science and Engineering, East China University of Science and Technology
Shanghai 200237, P.R. China
19001990@mail.ecust.edu.cn

Chen Shen
School of Information Science and Engineering, East China University of Science and Technology
Shanghai 200237, P.R. China
15221502172@163.com

***Abstract*—Automated Checkout (ACO) systems have high requirements for accuracy and speed in real retail scenarios. The research of ACO based on image recognition technology faces great challenges due to the lack of high-quality datasets, high merchandise granularity, and high model training costs. This project adopts PP-ShiTu algorithm based on mainbody detection, metric learning, and vector retrieval, and incorporates knowledge distillation and data enhancement strategies so as to carry out the implementation of merchandise recognition capability and effectively improve the prediction speed and recognition accuracy. With the feature learning dataset of retail scenes, the method adopted in this project can effectively balance the retrieval accuracy and speed, and improve the security and stability of the recognition process. Meanwhile, this project combines AIoT to connect "cloud, edge and end", forming an integrated and intelligent retail settlement platform.**

Keywords—Image Recognition; Metrics Learning; AIoT; Smart Settlement

I. INTRODUCTION

In the current "online + offline" new retail business model, the retail business has entered a period of accelerated development of intelligent and digital transformation and upgrading[1]. With the rapid development of computer vision technology and the concept of unmanned, automated and intelligent supermarket, the need for automatic product identification and automated checkout using image recognition technology and object detection technology has emerged, i.e. Automatic checkout (ACO) system[2]. The purpose of ACO is to automatically generate shopping lists based on images of products to be purchased, effectively reducing operational costs in the retail industry and improving the efficiency of merchandise settlement.

From the perspective of image recognition, ACO faces great challenges in terms of accuracy and speed. The number, variety and frequent updates of commodities make it difficult to collect a large number of commodity images reflecting real scenes, so the training of retail commodity models must solve the problem of small samples, and there is a lack of high-quality datasets, which makes the research difficult; in addition, the commodity fine-grained is high, and the visual differences between different commodities are usually small, so how to accurately distinguish commodities with high similarity is a big problem. At the same time, goods are updated very fast, and it would be very impractical to retrain the model once new products appear which would cost a lot. The recognition of retail products under vision solutions will also be affected by the product environment, such as shooting lighting, shooting clarity, and product placement, all of which will affect the accuracy of product recognition. Each of those factors has been considered to be challenging for automatic settlement research.

It is valuable to practice and study ACO projects. To address the above problems, firstly, a review of the current literature related to retail merchandise recognition is conducted, and then based on the ultimate pursuit of speed and accuracy in retail scenarios, the lightweight image recognition system PP-ShiTu is selected as the core algorithm of this paper to build an intelligent retail settlement platform. Experiments yielded an accuracy of 98.39% in the validation of a self-collected dataset with more than 250,000 items, and an inference speed of only 30ms for a single image under CPU, fully demonstrating the correctness of this retail settlement system.

II. RELATED WORK

The new retail model relies on the support of smart technology to identify and settle goods through smart devices. In this process, the accuracy and speed of recognition are critical to enhancing the user experience. There are three main technologies currently in use, scanning by barcode, recognition using radio frequency identification (RFID), and image recognition based on object detection.

Barcode identification technology[3, 13, 14] first appeared in the 1940s when two engineers, Joe Wood Land and Berny Silver, patented and put it into use in the United States. With the continuous development of barcode technology, various countries and regions in the world are now widely using barcodes for commodity identification and occupying an important position in automatic identification technology. Its working principle is to scan the barcode compiled according to

TABLE I. RELATED LITERATURE ON RETAIL IDENTIFICATION

Classification			Features
Barcode recognition			Product information can be accurately matched, but there are scanning problems and low efficiency[3].
Radio Frequency Identification			No manual scanning of products is required, but it is not possible to settle multiple products at the same time, which does not improve settlement efficiency[4].
Image recognition technology based on object detection	Monitor shopping behavior		Ready to take, but the equipment cost is large, the post-maintenance is difficult, the technology is complex and the error rate is high.
	Identification of settlement with the help of settlement equipment	Dataset	RP2K[5]、Product-10k[6]datasets are characterized by a small variety of goods and a current lack of high-quality datasets.
		Recognition Algorithm	(1) The traditional method of designing feature points manually has low accuracy and unsatisfactory application results. (2) Convolutional neural network-based graphics processing techniques. The following models and ideas are proposed for algorithm enhancement for fine-grained features： ①New hierarchical bilinear pooling framework[7]； ②Recurrent Attention Convolutional Neural Network RA-CNN[8]； ③Two concise bilinear representations[9]； ④Position-dependent depth metric (PDDM) unit[10]； ⑤Generalized iterative framework to learn low-dimensional features[11]； ⑥The objective function allows for joint comparisons between multiple negative samples[12].
		Model Training	Deep learning methods all require retraining models to adapt to new goods, with huge time and training costs.

certain rules, and the photoelectric converter converts the received reflected light of different strengths and weaknesses into corresponding electric signals and decodes the electric signals to obtain each information of the goods. Therefore, the barcode and goods on the market basically correspond to each other one by one to ensure the normal circulation of goods. At present, many supermarkets are using barcode technology for commodity settlement, through which commodity information can be accurately matched, but there are problems such as barcode wear and tear, variable position and difficulty in scanning, which affect the experience and efficiency of use.

With the improvement and application of radar, it has given birth to the radio frequency identification technology [4, 15-18]. At present, the theory of radio frequency identification continues to enrich and perfect, the standardization problem gets the attention, to a certain extent has improved the automation level. Its principle is to use the label and reads the electrostatic coupling between, receives and transmits the radio wave electronic label stores the commodity information, carries on the non-contact two-way communication to realize stores the information recognition and the data exchange. The merchant will affix the electronic label for each commodity, this method does not need to carry on the manual scanning to the commodity. However, since signals from multiple tags can interfere with each other, it is not possible to settle multiple items at the same time and improve settlement efficiency. Moreover, the electronic label price is dozens of times higher than the ordinary barcode label, and cannot be recycled. If the use of large quantities, the cost is too high, will largely reduce the enthusiasm of the market use. At the same time, the electronic label information is also easy to be illegally read and malicious tampering, there will be certain security risk.

In recent years, artificial intelligence has been developing rapidly, and one of the important branches, computer vision, has been emerging with new technologies and applications. There are two main forms of application of image recognition technology in the field of merchandise settlement, one is to monitor all the behaviors of customers in the shopping process for direct cloud settlement, and the other is to identify and settle the merchandise with the help of settlement equipment. The highlight of the first one is take-and-go, such as Amazon Go[19], which uses this approach. However, it requires the installation of a large number of hardware devices in physical supermarkets, including cameras, pressure sensors, infrared sensors, volume displacement sensors, light curtains, wireless networks, etc. The cost of the equipment is extremely high and the post-maintenance is troublesome. At the same time the technology needs to monitor the face, goods, behavior and other characteristics at the same time, which is complex and has a high error rate, making it difficult to promote. The second approach is the current mainstream form of image recognition technology application, such as Bingo Box, Convenience Bee and other companies are developing and using it. This method collects commodity image information with the help of settlement equipment, extracts commodity characteristics, identifies commodity classes and obtains information.

In the second form, there are several challenges to improve the accuracy and speed of product recognition: lack of high-quality datasets, difficulty in distinguishing similar products, high cost of model training with frequent product updates, and the product shooting environment affecting recognition accuracy. This project focuses on this approach.

Review of currently available datasets related to commodity identification: RPC[2] is a product dataset including single-product images captured in a controlled environment and multi-product images captured by the checkout system with different levels of annotation for the checkout images, the dataset contains 200 product categories and 83,739 images, but with relatively small product categories. Product-10k[6] contains nearly 10,000 items frequently purchased by online customers on JD.com. The

large-scale product labels are organized into a graph to show the complex hierarchy and interdependencies between products, and the dataset has a noise rate of less than 0.5% and contains nearly 150,000 images, but the distribution of images is highly uneven.

For the recognition accuracy problem, a traditional method is to manually design feature points to match the image recognition, Liu Jia et al. perform multi-resolution wavelet transform on the image, reconstruct the image approximate components, divide the feature point neighborhood area, establish a 32-dimensional feature point descriptor vector, use Euclidean distance to initially determine the matching points, and then use the integral image to further eliminate the false matching points, thus improving the matching accuracy[20]. Liu L. et al. proposed a simplified algorithm based on SIFT (SSIFT)[21]. A 12-dimensional vector based on a circular window is used instead of a 128-dimensional vector to effectively represent a feature point and improve the real-time performance of the algorithm. This method is not very accurate and is only suitable for studying small-scale images. However, the variety of commodities and the high accuracy requirement make the application not ideal. The deep learning field is growing rapidly, and convolutional neural network-based vision techniques have become a popular research direction, capable of automatically learning image features with better results than hand-designed feature points. One of the tasks of commodity recognition is classification, and methods of fine-grained image classification can be used to solve this problem. Y. Chaojian[7] et al. found that the interaction of local features between layers and the learning of fine-grained features are interrelated and also reinforce each other, thus proposing a new hierarchical bilinear pooling framework to integrate multiple cross-layer bilinear features and capturing the cross-layer interaction of local features to improve their representational power. J. Fu[8] et al. proposed a novel recurrent attentional convolutional neural network named RA-CNN for recursive learning of discriminative region attention and region-based feature representations by using mutual reinforcement. The learning performed at each scale size contains a classification subnetwork and an attention suggestion subnetwork, improving the relative accuracy, but it's more difficult to analyze at orders of magnitude from 100,000 to several million. Therefore, Y. Gao[9] et al. proposed two concise bilinear representations with the same discriminative power as the fully bilinear representation but with only a few thousand dimensions, and had experimentally been demonstrated that the effectiveness of the method in multi-dataset image classification and shot less learning. Also, metric learning can be used to solve such problems. Chen Huang[10] et al. proposed a position-dependent depth metric (PDDM) unit capable of learning a similarity metric that adapts to local feature structure, and local similarity-aware feature embedding not only showed faster convergence and higher performance on two complex image retrieval datasets but also brought superior generalization results in ensemble scenarios with its large margin property. To address problems of lack of training data, a large number of fine-grained classes, and high intra-class variance versus low inter-class variance, Y. Cui[11] et al. proposed a generalized iterative framework for fine-grained classification and dataset bootstrapping to learn low-dimensional features that embed anchor points on the stream shape of each class that captures intra-class variance and maintain distinctions between classes and demonstrated its validity. K. Sohn[12] et al. found that previous frameworks often suffer from slow convergence or even get stuck in locally optimal solutions, and proposed new metric learning to address this problem, with a proposed objective function that allows joint comparisons between multiple negative samples.

However, when the product categories are updated the object detection model of the above traditional deep learning methods need to be retrained to adapt to the new products, which spends lots of time and training costs and cannot meet the characteristics of the actual scene where the product categories updated very quickly. At the same time, there are usually tens of thousands of categories of goods in offline retail scenes, so it is difficult for the traditional object detection model to put all categories into the training set at one time, and the training is quite difficult.

In this project, we carry out the core development on the basis of the lightweight image recognition system PP-Shitu[22], and open source a dataset more suitable for retail feature learning scenarios, RPD357, based on open source large retail datasets such as RP2K[5] and Product-10k[6], which covers a total of 357 categories of goods commonly found in daily retail scenarios, with a total data volume more than 250,000. The image recognition algorithm is mainly composed of mainbody detection, feature extraction and vector retrieval. The algorithm uses various strategies to optimize each module in terms of backbone network selection and adjustment, loss function selection, data enhancement, learning rate transformation strategy, regularization parameter selection, pre-trained model use, and model cropping quantification; among them, the metric learning method with PP-LCNet[23] as the backbone of metric learning and ArcMargin as loss has improved the recognition accuracy, which can well solve the recognition problem of similar goods, and finally, the overall system recognition can be completed in 5ms on CPU. It not only solves the problem of the high similarity of goods and effectively improves the recognition accuracy, but also realizes the creative function of product category updating: it only needs to update the retrieval library to realize the rapid integration of new product categories without retraining the model. In this project, we integrate various data enhancement methods such as RandFlip, cutout[24] and AutoAugment[25] in the feature extraction model training process according to the actual scenario, so as to improve the model robustness. At the same time, AIOT is integrated with big data analysis technology, which together enables the collaborative empowerment of retail scenarios by the cloud and the edge, and realizes the application of deep learning algorithms.

III. ALGORITHM APPLICATION

The core of the intelligent settlement system is image recognition. Image recognition, that is, specifying a query image, the system is able to identify its category. The core of this project is the PP-ShiTu image recognition system consisting of mainbody detection, feature extraction, and vector retrieval[22]. Unlike other image recognition algorithms, PP-ShiTu only needs to update the corresponding retrieval library for unfamiliar categories to correctly identify the category of the query image without retraining the model, which greatly increases the usability of the image recognition system while reducing the need to update the model, perfectly satisfying the special multi-

category, small sample, high similarity and update mundane of goods in the traditional physical retail image recognition scenarios.

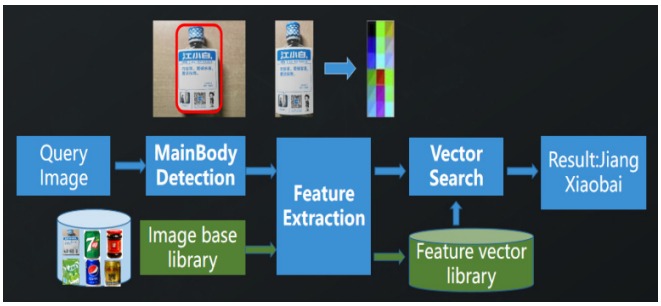

Figure 1. The structure of PP-ShiTu

A. Mainbody Detection

Mainbody detection technology is a detection technology that is currently applied and widely used. Specifically, it refers to detecting the coordinate positions of one or more subjects in a picture and then cropping down the corresponding areas in the image for recognition, so as to complete the whole recognition process. Mainbody detection is the pre-procedure step of the recognition task, which can effectively improve the recognition accuracy. Due to the requirement of real-time and high accuracy in real retail scenarios, this part of the project selects the lightweight mainbody detection model PicoDet[26], which incorporates optimized algorithms such as ATSS and Generalized Focal Loss, with a model size of only 30.1 MB, an mAP of up to 40.1%, and a single image prediction time (excluding preprocessing) of only 29.8ms.

B. Feature extraction

Feature extraction is a key part of image recognition, which is to transform the input image into a fixed dimensional feature vector for subsequent vector retrieval. Good features need to have similarity preservation, i.e., in the feature space, images with high similarity are more similar to their features and images with low similarity being less similar to their features. Deep metric learning is usually applied in this direction.

Metric learning[27] is a form of machine learning that automatically constructs a task-specific metric function based on training data, intending to learn a transformation function that maps data points from the original vector space to a new vector space where similar points are closer together and non-similar points are further apart, making the metric more task-appropriate.

Due to the extreme pursuit of accuracy and speed in the retail industry, PP-LCNet-x2-5, which is optimized for Intel CPU and supports MKLDNN acceleration, is used as the backbone network in this project. The network uses H-Swish as the activation function, adds an SE module near the tail of the network to improve the accuracy, and places a larger convolutional kernel in the middle and rear of the network to increase the inference speed. Compared with other lightweight SOTA models, this backbone network is able to further improve the model performance without increasing the inference time, and eventually significantly outperform the existing SOTA models. In addition, Linear Layer is chosen for the Neck part, ArcMargin for the Head part, and CELoss for the Loss part, fusing the Cutout image augmentation method, which eventually achieves an accuracy of 98.39%[28] and an inference speed of only 30ms in the validation of the self-collected dataset.

C. Vector search

Vector retrieval is widely used in image recognition and image retrieval. Its main objective is to obtain a similarity ranking for a given query vector by performing a feature vector similarity or distance calculation with all the vectors to be queried, in the already established vector library. In this project, the HNSW32[29] graph indexing method of Faiss is selected to be used. This method can effectively balance the retrieval accuracy and speed.

D. Security Analysis

The recognition of retail products under vision solutions will be affected by the environment factors, such as shooting lighting, shooting clarity, and product placement. Among them, to reduce the influence of the shooting environment on the model recognition, various data enhancement methods such as RandFlip, cutout[24], and AutoAugment[25] are integrated in the training process of the feature extraction model based on the actual scene, so as to improve the model robustness. In addition, light-assisted devices can be added to help shooting in the actual operation process.

Figure 2. Shooting light differences

Figure 3. Partial and full overlap of products

To address the problem of missing recognition of goods

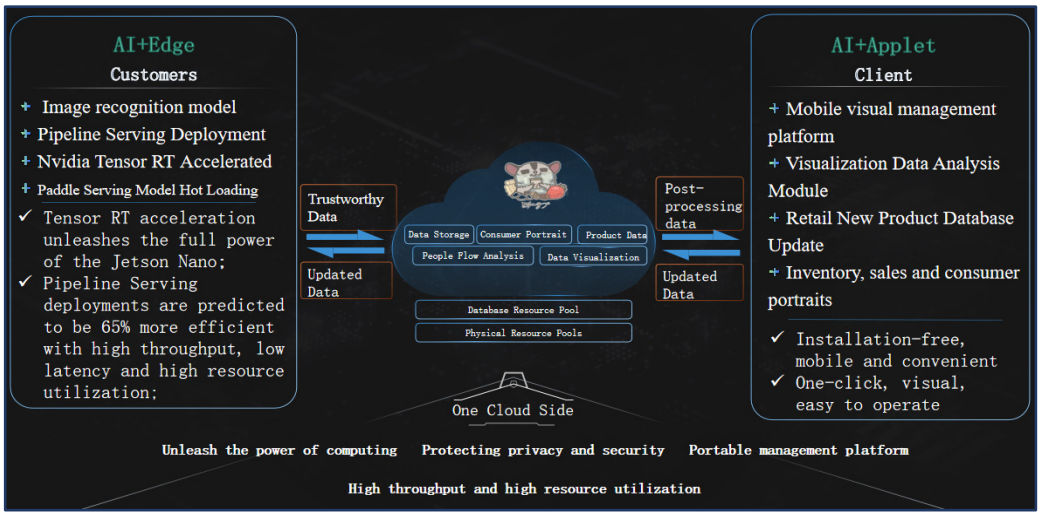

Figure 4. System Architecture

caused by different angles of object placement, in the practical application of this project, images of goods with different angles can be provided in the retrieval library for vector search to improve matching accuracy. As for the problem of missing recognition of products caused by overlapping products, the method proposed still performs well after field tests under the condition that a product is not completely obscured by other products. If it is completely obscured, the overall recognition security will be subsequently improved by adding infrared perception.

IV. System Design

This project builds the main architecture of the system with “machine vision” and “intelligent hardware” as the technical core, supplemented by mobile applications for intelligent management and data analysis of goods.

A. Architecture Design

Based on the actual traditional retail settlement application requirements, the AIoT technology is integrated with cloud-side integrated smart hardware device to photograph the purchased goods as a whole, adopts the image recognition technology based on mainbody detection, feature extraction and vector retrieval, and makes prediction through the Pipeline service deployment method that takes into account throughput and efficiency, so as to accurately locate and identify the purchased goods. The information of prediction result is linked with the cloud database to obtain the unit price and total price of each product, and the price list is returned to the end side for display, and the customer pays according to the price list. The cloud database updates the inventory and sales of each product in real-time according to the products sold and feeds back to the management applet for real-time display. The management applet is equipped with functions such as product category update, product information modification, and product data analysis.

An industrial-level end-to-end intelligent retail settlement system is proposed to help large offline retail stores improve the efficiency of retail settlement, provide customers with a frictionless retail experience, reduce actual operating costs, truly achieve "cost reduction and efficiency increase", empower the unmanned, automated and intelligent level of the entire physical retail industry, and help traditional retail transform into a new retail form of "consumer experience-centered, data-driven pan-retail".

B. Cloud-edge collaboration

Cloud-edge collaboration mainly refers to the complementary and synergistic relationship between edge computing and cloud computing in application scenarios[30]. In fact, edge computing is mainly responsible for real-time, short-period data processing tasks and real-time processing and execution of local business, providing high-value data for the cloud. Cloud computing has massive scalable storage capacity to process and analyze data that is non-real-time and relatively long-cycles. From the current practical situation faced by industry, cloud computing and edge computing need to be closely combined to meet the needs of the scenario more highly.

In view of the ultimate requirements of low latency and high throughput in the physical retail settlement scenario, the edge side adopts the Python Pipeline serving service-oriented deployment method that balances throughput and efficiency and supports single-operator multi-model combination scenarios and asynchronous mode, fully releasing the AI computing capability of the edge devices, with an average overall process time of about 1 second and a 65% improvement in prediction performance compared with the traditional method. The data obtained at the edge communicates with the cloud and relies on the cloud's massive scalable storage capacity for storage, while real-time data processing and analysis are performed, and the analysis results are synchronized with the large data visualization screen.

Considering the characteristics of blocky commodity category update and high real-time information loading of physical retail, taking full advantage of the lightweight and

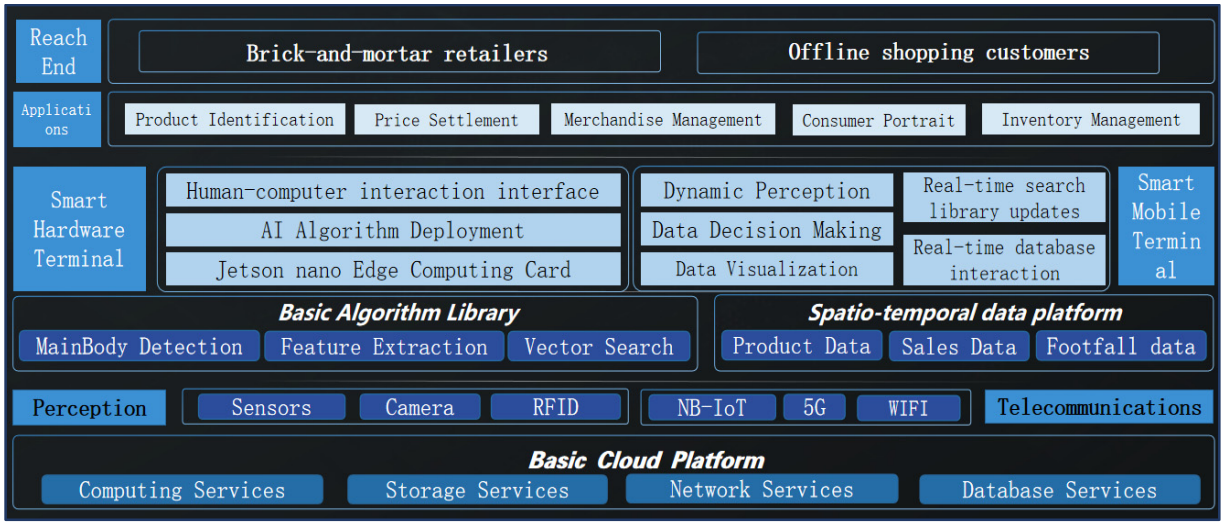

Figure 5. Application Architecture and Capability System

convenient features of mobile WeChat applets, it provides the function of commodity category update and information modification through mutual trust with the cloud, which is convenient for entity retail operators to update merchandise information anytime and anywhere.

C. *Deploying applications*

At the smart hardware terminals, the C/S platform is used for development. The C/S architecture allows the smart hardware terminals to be directly connected to the server, which makes the endpoint-to-endpoint model more secure and reduces communication traffic due to the direct connection. In addition, thanks to the advantage of the cloud-side collaboration of smart hardware terminals, logical computing tasks can be shared for the server for data processing and data transit, thus simplifying the complex task flow to obtain a fast response from the cloud. Meanwhile, the C/S platform ensures the security of data information during cloud-side communication and can make full use of local hardware facilities to handle some logical operations, avoiding the waste of resources.

(a) (b)

(c) (d)

Figure 6. Deployment Application Effect

In the intelligent mobile terminal, the WeChat applet platform is adopted. Relying on WeChat's large platform of one billion traffic, WeChat applets are uniquely advantageous for the speed of service acquisition and the efficiency of management. In addition, WeChat also provides numerous interfaces to the applets, and developers can use these capabilities to efficiently develop the required functions, effectively shortening the system development cycle and reducing the complexity of system maintenance and iteration.

V. Experimental Results

Considering the concerns of retail scenarios on product data, model accuracy, prediction speed and model robustness, the model is trained and tested with self-collected datasets, using PicoDet as the main detection model and PPLCNET-x2-5 as the feature extraction backbone, incorporating knowledge distillation and data augmentation strategies to improve model robustness and accuracy.

TABLE II. RESULTS OF ABLATION EXPERIMENTS

Model	use ssld	num_epoch	batch_size/GPU cards	base_lr	use cutout	top1 recall	Intel-Xeon-Gold-6148 time(ms) bs=1
PP_LCNet_x2_5	N	400	256/4	0.01	N	98.19%	29.595
PP_LCNet_x2_5	N	400	256/4	0.01	Y	98.21%	29.595
PP_LCNet_x2_5	Y	400	64/4	0.002	N	98.38%	29.595
PP_LCNet_x2_5	Y	400	64/4	0.002	Y	98.39%	29.595

Notes：

1. The training machine for this experiment is Tesla V100；
2. Evaluation metric: Recall (recall) - indicates the number of predicted positive cases with positive labels / the number of cases with positive labels.

A. Retail scenario feature learning dataset

In this project, we open source a dataset that is more suitable for retail feature learning scenarios based on retail application scenarios with reference to large retail datasets such as RP2K and Product-10k that are currently open source. The dataset covers 357 categories of common products in daily retail scenarios, with a total data volume of over 250,000, and a 7:3 division between the training and testing sets. Most of the image data in this dataset are collected from real scenes, and the unbalanced distribution problem is avoided from the data source. In addition, the dataset contains fine-grained products such as "Lactobacillus Original, Lactobacillus Mango", which can better simulate the product data of actual retail scenes.

B. Knowledge distillation

Knowledge distillation[31], is one of the mainstream methods for compression of models at present. In recent years, deep neural networks have had significant breakthroughs in natural language processing, computer vision, and other fields. In particular, deep neural networks in the context of big data, increasing their number of parameters by means of reasonable construction of network models can significantly improve model performance, but also make the model complexity extremely high. Therefore, large models face the problem of high usage costs in the usual practical application scenarios. In contrast, knowledge distillation specifically refers to the use of a teacher model to guide a student model to learn a specific task, ensuring that the small model gets a large performance improvement with the same number of parameters.

C. Ablation experiments

In this project, based on the actual retail application context, the SSLD knowledge distillation scheme is adopted, and the following ablation experiments are conducted in the context of the self-collected dataset by adding automatic broadening and Cutout to the data augmentation during the training process：In summary, the model can achieve more satisfactory results in the test set after each parameter adjustment, data enhancement, knowledge distillation and other strategies, which meet the relevant requirements for model accuracy and prediction speed in the retail scenario.

VI. CONCLUSION

This project provides insight into the problems of physical retail enterprises, breaks the traditional object detection algorithm applied to the intelligent retail industry with large cost and difficult training, adopts an innovative image recognition algorithm based on mainbody detection, feature extraction and vector retrieval, fully considers the ultimate pursuit of speed, accuracy and data security in retail scenarios, and integrates various data enhancement algorithms to achieve the optimal balance effect in terms of both accuracy and speed, providing a reference example for unmanned retail visualization intelligence in the new retail industry to solve the painful problems of special image recognition scenarios with multiple categories, small samples, high similarity and frequent updates.

REFERENCES

[1] T. Zhang, "Research on Digital Transformation of Traditional Supermarkets under New Retailing Format - A Longitudinal Single Case Study Based on RT-Mart (in Chinese),"in *Management and Technology of SME*, no. 11, pp. 103-105, 2021.

[2] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249,* 2019.

[3] J. Li and S. Zhan, "The development and application of barcode technology (in Chinese),"in *Computer and Digital Engineering*, vol. 37, no. 12, pp. 115-118+154, 2009.

[4] Q. He and L. Xia, "Wireless radio frequency identification technology application overview (in Chinese)," in *Modern Architecture Electric*, vol. 2, no. 08, pp. 1-4, 2011, doi: 10.16618/j.cnki.1674-8417.2011.08.003.

[5] J. Peng, C. Xiao, and Y. Li, "RP2K: A large-scale retail product dataset for fine-grained image classification," *arXiv preprint arXiv:2006.12634,* 2020.

[6] Y. Bai, Y. Chen, W. Yu, L. Wang, and W. Zhang, "Products-10k: A large-scale product recognition dataset," *arXiv preprint arXiv:2008.10545,* 2020.

[7] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 574-589.

[8] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438-4446.

[9] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317-326.

[10] C. Huang, C. C. Loy, and X. Tang, "Local similarity-aware deep feature embedding," *Advances in neural information processing systems,* vol. 29, 2016.

[11] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1153-1162.

[12] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems,* vol. 29, 2016.

[13] H. Song, "Identification of commodity barcode (in Chinese)," in *Chinese Brand and Anti-counterfeiting*, no. 03, pp. 76-77, 2006.

[14] Y. Wang and Z. Dou, "Research on image-based barcode reader (in Chinese),"in *Video Engineering*, no. 01, pp. 89-91, 2007, doi: 10.16280/j.videoe.2007.01.026.

[15] Y. Wu, "Radio frequency identification (RFID) technology research status and development prospects (in Chinese)," in *Microcomputer Information*, no. 32, pp. 234-236+230, 2006.

[16] X. Zhang, M. Zheng, and Z. Xing, "New commodity identification technology - radio frequency identification technology (in Chinese),"in *Guangdong Print*, no. 04, pp. 54-55, 2005.

[17] L. Wang, " Research on RFID-based algorithm for analyzing supermarket shopping data (in Chinese)," Master Thesis. Shanxi: Taiyuan University of Technology, 2017.

[18] X. Chen, "Wireless radio frequency identification (RFID) technology development overview (in Chinese)," in *Information technology & Standardization*, no. 07, pp. 20-24, 2005.

[19] H. Gu, "From "Amazon Go" to see the realization of unmanned supermarket in the era of artificial intelligence (in Chinese)," in *Oigital Communication World*, no. 03, pp. 151-152+154, 2017.

[20] J. Liu, W. Fu, W. Wang, and N. Li, "Image matching based on improved SIFT algorithm (in Chinese)," in *YiQi YiBiao Xue Bao Chinese Journal of Scientific Instrument*, vol. 34, no. 05, pp. 1107-1112, 2013, doi: 10.19650/j.cnki.cjsi.2013.05.022.

[21] L. Liu, F. Peng, K. Zhao, and Y. Wan, "Fast image matching using simplified SIFT algorithm (in Chinese)," in *Infrared and Laser Engineering*, no. 01, pp. 181-184, 2008.

[22] S. Wei *et al.*, "PP-ShiTu: A Practical Lightweight Image Recognition System," *arXiv preprint arXiv:2111.00775,* 2021.

[23] C. Cui *et al.*, "PP-LCNet: A Lightweight CPU Convolutional Neural Network," *arXiv preprint arXiv:2109.15099,* 2021.

[24] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552,* 2017.

[25] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113-123.

[26] G. Yu *et al.*, "PP-PicoDet: A Better Real-Time Object Detector on Mobile Devices," *arXiv preprint arXiv:2111.00902,* 2021.

[27] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry,* vol. 11, no. 9, p. 1066, 2019.

[28] X. Yan. "Smart Container." https://github.com/thomas-yanxin/Smart_container, 2022.

[29] K. Echihabi, "High-dimensional vector similarity search: from time series to deep network embeddings," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 2829-2832.

[30] E. Xu and E. Dong, "Exploration and Practice of Synergistic Development of Cloud Computing and Edge Computing (in Chinese)," in *Communications World*, no. 09, pp. 46-47, 2019, doi: 10.13571/j.cnki.cww.2019.09.024.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531,* vol. 2, no. 7, 2015.