IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 January 2024, accepted 30 January 2024, date of publication 2 February 2024, date of current version 13 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3361675

RESEARCH ARTICLE

BP-YOLO: A Real-Time Product Detection and Shopping Behaviors Recognition Model for Intelligent Unmanned Vending Machine

JINGXIANG LI[1], FUQUAN TANG[1,2], CHAO ZHU[1], SHIWEI HE[3], SHUJIN ZHANG[4], AND YU SU[1]

[1]School of Geomatics, Xi'an University of Science and Technology, Xi'an, Shaanxi 710054, China
[2]Key Laboratory of Coal Resources Exploration and Comprehensive Utilization, Ministry of Natural Resources, Xi'an 710021, China
[3]Courant Institute of Mathematical Science, New York University, New York, NY 11201, USA
[4]School of Mechanical and Electronic Engineering, Northwest A&F University, Xianyang, Shaanxi 712100, China

Corresponding author: Fuquan Tang (2504557922@qq.com)

This work was supported in part by the National Science Foundation of China under Grant 52274168, and in part by the Shaanxi Natural Science Foundation of China under Grant 2016JM5048.

ABSTRACT Intelligent unmanned vending machines (UVMs) based on machine vision have attracted great attention in the unmanned retail industry. However, due to the complexity of practical application scenarios and environments, the existing vision-based intelligent UVMs face challenges related to missed-detection and mis-detection of product, and require costly physical components such as the infrared radio frequency sensors to capture shopping behaviors. In this study, we propose a BP-YOLO, the real-time model that integrates optimized YOLOv7 and BlazePose for product detection and shopping behaviors recognition. BP-YOLO can accurately detect the products purchased by consumers and their shopping behaviors in complex scenarios. To address the problems of missed-detection and mis-detection, we introduce the 3D attention mechanism SimAM and the deformable ConvNets v2 (DCNv2) to recombine and optimize the one-stage object detection model YOLOv7. This method reduces the interference of the invalid information in complex scenarios by adaptively weighting each channel and 3D spatial features, focuses on feature information in a sparse space, and minimizes the loss of feature information during the transmission process based on multi-scale feature extraction and fusion. To recognize and judge the shopping behaviors of consumers, we track the hand and arm key points of consumers using the pose estimation model BlazePose. Using the mAP@[0.5:0.95] as the evaluation metric for product detection, the experimental results on a customized product dataset show that BP-YOLO achieves an average accuracy of 96.17% for all product categories detection; the average success rate of consumer shopping recognition reaches 92%, 98%, and 94.7% under three light and noise intensity, respectively. Therefore, our BP-YOLO model for intelligent UVMs has effectiveness in commercial deployment.

INDEX TERMS BP-YOLO, BlazePose, product detection, shopping behaviors recognition, unmanned vending machines.

I. INTRODUCTION

With the development of the Internet of Things, big data, and artificial intelligence, the concept of 'intelligent vending' has emerged in the unmanned retail industry. This broke the limitations of traditional mechanical unmanned vending machines (UVMs), such as high product loss rate and inconvenient management. The integration of "artificial intelligence", "machine vision", and "UVM" resulted in two types of intelligent UVM based on static vision [1] and dynamic vision [2].

The static vision-based UVM uses multiple cameras inside the cabinet to compare the remaining products on each shelf before and after the cabinet door is opened, and settles the

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

products purchased by the consumers. This method requires high camera hardware costs and strict product placement constraints, as stacked or obscured products can lead to missed-detection. Dynamic vision-based UVM records the process of consumers picking up products through cameras inside the cabinet, and uploads them to the cloud where an object detection algorithm recognizes the product categories from the video frames. In product detection, poor shopping environments or light inside the cabinet can cause overexposure, distortion and excessive noise in the captured image or video frames. Moreover, consumer shopping behaviors can cause partial occlusion and deformation of product packaging. These issues can result in the missed-detection and mis-detection of product. In the application of judging the consumer's purchasing behavior, the static and dynamic UVMs rely on optical physical instruments such as red sensors on the door frame to sense the consumer's hands movements, or cooperate with radio frequency identification (RFID) technology [3] and gravity sensor technology [4], [5]. While the technology provides some convenience in application realization, it greatly increases production costs and mechanical design complexity. The current vision-based intelligent vending cabinet provides a new solution by selecting a certain number of valid frames from the video of consumer shopping, and tracks and compares the products in these frames to determine their purchasing behaviors. However, this method has high requirements for the robustness, hardware and computing performance of the algorithm, which makes it difficult to achieve real-time and accurate judgments.

To address the challenges posed by complex shopping scenarios inside the UVM cabinets and achieve shopping behaviors recognition while reducing the dependence on mechanical tools and sensors, we propose BP-YOLO, a novel unmanned vending model. This model can maintain real-time and accurate recognition abilities for various products and shopping behaviors in complex shopping environments. In product detection, BP-YOLO integrates the SimAM, a 3D attention mechanism, and the DCNv2, a deformable convolutional network, on top of YOLOv7, a one-stage object detection algorithm. First, the objective of SimAM is to reduce the interference of extraneous irrelevant information in intricate shopping environments and scenarios, and enhance the model's perception and acquisition of the target product's feature information. Second, DCNv2 can readjust the extension of attention in space and channel, as well as the spatial-channel dependency of convolution kernels, improving the focusing ability of convolution kernels on discrete feature information and reconstructing the feature information of product that have wrinkles, deformation or partial occlusion. These modules help the model to effectively locate and capture the valid feature information of product packaging and reduce the probability of mis-detection and missed-detection. In shopping behaviors recognition, BP-YOLO incorporates BlazePose, a pose estimation model, to recognize consumers shopping behavior based on vision. This method can replace traditional methods such as multi-frame comparison, weighing, RFID and optical sensors, which can reduce the production cost and mechanical design complexity of intelligent UVMs. BlazePose can track and locate consumers' palms, wrists and arm positions, and define shopping behavior categories by angles formed by key points. It can further judge consumers' purchase behavior during shopping process, and determine whether consumers successfully purchased a certain product. This improves shopping efficiency, reduces production cost, and provides consumers with the convenient shopping experience of "take it away".

We evaluate BP-YOLO on a simulated real UVM shopping scenario, where we define a platform for placing products. And BP-YOLO is applied to product detection and shopping recognition tasks. The main contributions of this study are as follows:

1) In the task of product detection, BP-YOLO can achieve real-time and accurate recognition of various products in complex shopping environments.
2) BP-YOLO provides an effective method to judge consumer shopping behaviors.
3) For new products, BP-YOLO can adapt to new products through cloud services, which avoids the problem of cumbersome products update of traditional UVMs.

II. RELATED WORK

Machine vision is the key component of the intelligent UVM retail algorithms. With the rapid development of high-performance graphics processing units (GPUs), the cloud service deployment of real-time detection based on visual deep learning models has become more feasible. However, in the domain of vision-based intelligent UVMs, the product detection faces challenges such as product deformation, partial occlusion and noise, and consumer shopping behavior recognition also requires improvement. Therefore, reducing mis-detection and missed-detection in product detection tasks, and efficiently and low-costly judging consumer purchase behavior are the research hotspots in this field. In this section, we will review the related work on object detection models, behavior recognition techniques and intelligent UVM retail algorithms.

A. OBJECT DETECTION

The key to recognize the product category in intelligent vision-based UVMs is to apply object detection algorithms, which can be categorized into two types: one-stage and two-stage. Two-stage algorithms divide the object detection process into two steps: first, they generate candidate regions that may contain objects; second, they classify the candidate regions and refine their bounding boxes. This type of algorithms achieves high prediction accuracy, but suffers from low speed and is not appliable to real-time object detection tasks, such as R-CNN [6], Fast R-CNN [7], and Faster R-CNN [8]. One-stage algorithms skip the candidate region generation step and directly output the final localization and classification results in one step. This type of algorithms has fast computation speed and can meet real-time requirements such as SSD (Single Shot Multibox Detector) [9] and YOLO (You Only Look Once). However, SSD has low detection accuracy for small-sized or occluded objects and lacks the ability to extract feature information from objects. Its detection efficiency is comparable to that of Faster R-CNN [10]. The YOLO network framework design has evolved from v1 to v7, focusing on improving the speed, accuracy, and recall of object prediction. In industrial detection tasks, YOLO has more efficient feature learning ability and outstanding generalization performance compared to other detection models. Therefore, this study adopts YOLO as the model basic framework.

YOLO [11], [13] is a classic one-stage object detection model with an end-to-end structure in deep neural networks. It can reduce the feature reuse rate and the redundant feature computation amount. It is mainly applied in industrial detection tasks that require high efficiency. Yin et al. [14] proposed to apply YOLOv4-Tiny in the security detection of UVMs, achieving the recognition of consumer damage behavior for the products and the machine. Horng and Huang [15] and Liu et al. [16] used YOLOv4 and YOLOv3-tinyE respectively in the product detection task of intelligent UVMs. However, they did not fully consider the influence factors of intricate real shopping environment, such as partial occlusion, deformation, damage, and light of the product packaging in the real vending cabinets, which affected the product detection accuracy. Compared with other versions of YOLO framework, YOLOv7 proposed by Wang et al. [13] focuses more on the optimization of modules, network architecture, and training process. It improves the accuracy and speed of real-time forward inference in object detection tasks without increasing the cost consumption of parameter computation. Zhu et al. [17] used YOLOv7 to accurately extract the classification and localization information of sand belt surface defects in the sand belt grinding surface defect classification task. In the comparative experiment with other models, YOLOv7's mAP and frame rate reached 0.907 and 83fps respectively, which were much higher than the performance of other object detection models.

B. BEHAVIOR RECOGNITION

Behavior posture estimation networks that track human key points through visual sensors have become a heated topic of current research, due to the innovative development of many new lightweight networks in the vision field. Bazarevsky et al. [18] proposed a human key point tracking detection algorithm based on lightweight convolutional neural network called BlazePose, which is used for human skeleton generation and pose estimation. Ke et al. [19] used BlazePose algorithm to analyze the angles in human finger movements, and built a gesture action classification model to achieve the control function of human-computer interaction. The model achieved an accuracy of 87% for the recognition of four gesture commands. To enable the wide application of BlazePose in vision tasks, Liu et al. [20] integrated BlazePose with other deep learning models, and proposed a sailor fall automatic detection algorithm based on BlazePose-LSTM. The algorithm achieved a detection accuracy of 98.5% for the sailor fall behavior, and a detection frame rate of 29fps on CPU, which cannot meet the real-time detection requirements. Huang et al. [21] proposed an object-based sequential action recognition technology based on integrating BlazePose with LSTM and improved YOLOv4, which normalized and supervision of people's series of actions in activities, and verified that BlazePose had a good generalization ability in specific complex environments. In the research of consumer shopping behavior recognition in UVM cabinets, Horng and Huang [15] proposed to use OpenPose [22] human pose estimation algorithm to recognize consumer behavior in the unmanned store recognition system, preventing shop products from being stolen. Due to low frame rate and generalization of OpenPose in tracking human key points under CPU, this study employed BlazePose to track consumer palm, wrist and arm skeleton points in consumer shopping behavior recognition task, constructed specific behavior classification corresponding to angles, and achieved the recognition of shopping behavior. The main method is discussed in detail in the third part.

C. INTELLIGENT UVMS RETAIL ALGORITHM

Many scholars have conducted a sheer volume of research and achieved some accomplishments in product dataset, product detection and UVMs retail algorithm tasks.

For the existing product datasets, Microsoft created the Grocery Store DataSet [23], which mainly consists of 5125 images taken from supermarket shelves, with an average of about 5 product instances per image. Zhang et al. [24] constructed a large-scale multi-category beverage detection benchmark dataset based on vision-based UVMs with the real shopping environments, which mainly covers 10 common beverages in the retail market, with an average of about 4 categories of products per image. However, the images only describe the top structure contour information of the beverages. Wei et al. [25] constructed a high-quality dataset that contains multi-angle feature information of product instead of a single top contour information. Chen et al. [26] proposed a method to construct an excellent product dataset, introducing lighting changes and multi-angle shooting technologies to enhance the generality and diversity of the product dataset. In view of the deployment environment and scenario of UVMs, this study collected 15 categories of products, including various beverages, biscuits and chips and other common products. To simulate the application scenario environment of real UVMs, 15 categories of products were shoot from multiple angles, and image augment technologies were used to produce product image datasets containing different intensities of light and noise, enriching the feature information of product under various scenarios. In order to reduce the impact of the model on the product recognition accuracy, where consumers hand may cause partial occlusion of the product packaging during the shopping, some datasets were partially occluded by using hands to cover the product packaging during shooting.

In the research field of smart unmanned vending machine product detection methods and related retail algorithms, Sun et al. [27] proposed a Template product detection system that segmented multiple layers of products on unmanned shelves into single products and identified each product according to the segmentation results. Xu et al. [2] proposed a smart unmanned vending machine algorithm based on the Internet of Things and artificial intelligence, using a single-stage SSD target detection algorithm to achieve a static detection accuracy of 91.79% for products. Liu et al. [16] first collected product images inside the unmanned vending machine by using a binocular camera, then used Harris [28], [29] method to extract and match product features from the images, solved image distortion by feature point stitching and fusion, and finally used a single-stage YOLOv3-tinyE [12] to perform target detection on the products in the images. Xia et al. [4] combined product recognition and weighing methods to achieve a retail solution that integrates product weighing, recognition, and online settlement without RFID. Based on these studies, Xu et al. [24] divided the vision-based smart vending machine recognition strategy into static and dynamic recognition processes. In the static recognition task, they identified the products that consumers dynamically took and judged whether consumers purchased a certain product. In the dynamic recognition task, they used an infrared laser sensor to determine whether consumers' hands left the vending machine and used physical hardware to determine whether consumers purchased a certain product. In addition, to enable the target detection model to adapt to real shopping scenarios and maintain efficient recognition performance for damaged or deformed products, Li et al. [30] designed a DrtNet product detection model for smart unmanned vending machine scenarios. Its backbone structure adopts deformable convolution kernels and uses infrared sensors to determine consumer purchase behavior and proves its feasibility and effectiveness. Due to the variety of shapes and sizes of products in unmanned vending machines, complex deployment environments, and easy occurrence of missed detection and false detection situations, existing smart vending product detection algorithms have not provided corresponding solutions for this problem. In addition, many studies only provide one infrared sensor method for consumer shopping behavior judgment. Therefore, existing methods have certain constraints on the development and cost of smart vision vending algorithms. This study proposes BP-YOLO model to detect product categories and identify consumer shopping behavior in real time. The next section will elaborate on BP-YOLO model in detail.

III. METHOD

A. OVERVIEW

This study proposes a vision-based intelligent unmanned retail model that consists of two components: product detection and shopping recognition, based on the BP-YOLO method.

1) For product detection, we propose to optimize the performance of the detection by introducing a 3D attention mechanism and a deformable convolution module into the backbone network and the feature pyramid network (FPN) of the BP-YOLO framework, respectively. We also reorganize the network and the related modules, and optimize the loss function.
2) For shopping recognition, we use a visual fusion model that combines skeleton point generation and pose estimation to track the shopping behavior of the consumers. We then apply BP-YOLO to dynamically recognize the products picked by consumers and their shopping behavior in real time.

B. PRODUCT DETECTION

In the real shopping environment, the camera capture of products will be affected by various external factors that are inevitable and complex. For example, during the process of consumers picking up products, the product packaging may be squeezed or occluded by their hand; if the product packaging color is too dark, and the surface is smooth, the indoor or outdoor lighting will cause light reflection on the product packaging surface, resulting in overexposure, noise or distortion of captured products images. These factors will affect the product detection, resulting in mis-detection or missed-detection of product. This section proposes an effective object detection scheme to solve the above interferences, mainly including adding the SimAM attention mechanism in the Backbone layer of the original YOLOv7 framework to enhance the feature representation, replacing the conventional convolution in the CBS module that implements cascading in the FPN [31] structure with the deformable convolution to adapt to the shape variations of products, and using a Focal Loss function based on cross entropy to reduce the class imbalance problem.

The YOLOv7 object detection framework consists of three components: Input, Backbone, and Head. The Input component preprocesses the input image by applying image augmentation, size alignment, and stitching methods. The Backbone component performs feature extraction at different scales by using downsampling operations. The Head component performs feature inference and outputs the detection results. The Backbone component is composed of multiple CBS, ELAN, and MPConv modules; the Head component is composed of CBS, SPPCSPC, UpSampling, and ELAN modules. To achieve multi-scale feature fusion for object detection tasks, YOLOv7 constructs a FPN by cascading modules at different layers of Backbone and Head, as shown in FPN (the left part of Figure 1).

FPN uses CBS modules with 1×1 convolution kernels to build a lateral connection structure, which merges the multi-scale feature maps that capture different semantic information after hierarchical convolution of the input

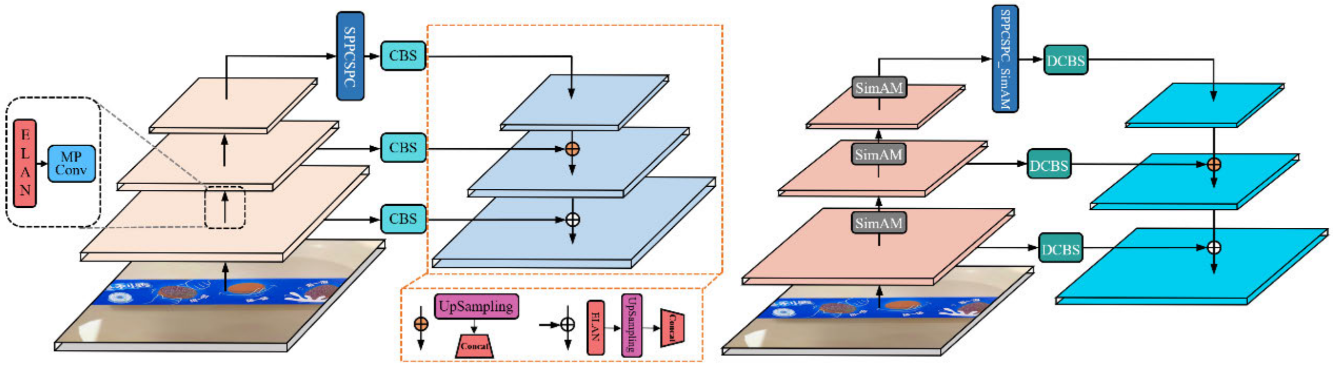


FIGURE 1. Overall structure design of FPN (left) and optimized FPN (right). The optimized FPN gradually extracts and fuses pyramid features by adding SimAM attention mechanism and DCNv2 module. For simplicity, only three-layer feature cascading structure is drawn here.

image. The high-level features have lower resolutions but stronger semantic information, while the low-level features have higher resolution but weaker semantic information. The Backbone component contains multiple CBS and ELAN modules that are connected by batch normalization and SiLU activation function skip connections, which are responsible for reducing the resolution and extracting the low-level features of the input high-resolution feature map. After passing through the ELAN and SPPCSPC modules in the FPN structure, three feature map sequences of $80 \times 80 \times 512$, $40 \times 40 \times 512$, and $20 \times 20 \times 512$ are obtained from different depths of the network, respectively. Their resolution decreases progressively, and their semantic information increases progressively. Next, three feature map sequences with reduced channel dimensions are obtained by using CBS modules with 1×1 convolution kernels. The lowest-resolution feature map is upsampled by nearest neighbor UpSampling to increase its resolution and then concatenated with the previous layer feature map in the sequence, which achieves the effect of obtaining low-resolution semantic information in the high-resolution feature map and passing it to the low-resolution feature map for feature reuse. This step is repeated to generate a multi-scale feature pyramid. This method mainly compensates for the lack of semantic information in low-level features and the low resolution of high-level features, and achieves the effect of mapping and efficiently fusing features at different levels. However, the process of passing feature information from high-level to low-level and using CBS modules with 1×1 convolution kernels to control channel dimension for multi-dimensional feature concatenation will cause significant semantic information loss due to the drastic change of channel dimension, and the direct fusion of multi-dimensional features ignores the semantic gap between feature maps at different depths of the network [32].

To address the aforementioned problems, this study proposes a novel model that incorporates visual attention mechanism, which enables the model to focus on the important feature semantic information and assign high weights, while ignoring the irrelevant feature semantic information and assign them low weights. The 3D attention mechanism SimAM, compared with the traditional spatial and channel attention mechanisms, mainly evaluates the unified weight values of the three-dimensional feature weights that each channel and spatial neuron pays attention to in the feature extraction process, enhances the expression of the important feature information in the object anchor box, and weakens the interference of the background and surrounding complex scene information. In addition, the relevant operators in SimAM are selected by the expression of the energy function, omitting the introduction of learnable parameters, which does not hinder the computational efficiency of the model after adding this module. In this study, SimAM attention mechanism module is added to the ELAN module in the Backbone layer, and SPPCSPC module in FPN structure is replaced by SPPCSPC_SimAM module (as shown in Figure 1 and 2). This method guides the aggregation of key features by SimAM in the feature extraction and propagation process, improves the model's perception of key feature information of the object, reduces the loss of high-level information in the transmission process, and improves multi-scale feature learning. Secondly, to make the model more attentive to sparse spatial information in tasks, DCBS module is introduced, which differs from CBS module in that it uses deformable convolution DCNv2 to implement convolution operation. DCBS module establishes lateral connection between Backbone and Head layers in FPN, and deformable convolution expands the receptive field of convolution kernel, mapping each layer's connected information to Backbone. Considering the irregular shape of product packaging, product packaging deformation and partial occlusion, CBS module in Backbone is replaced by DCBS module, which improves the model's flexible extraction ability of effective features in input images. When DCBS with deformable convolution kernel is applied to FPN cascading module lateral splicing, it effectively enhances feature information expression from different network depths and improves semantic gap [33], as shown in FPN (the right part of Figure 1) in Figure 1.

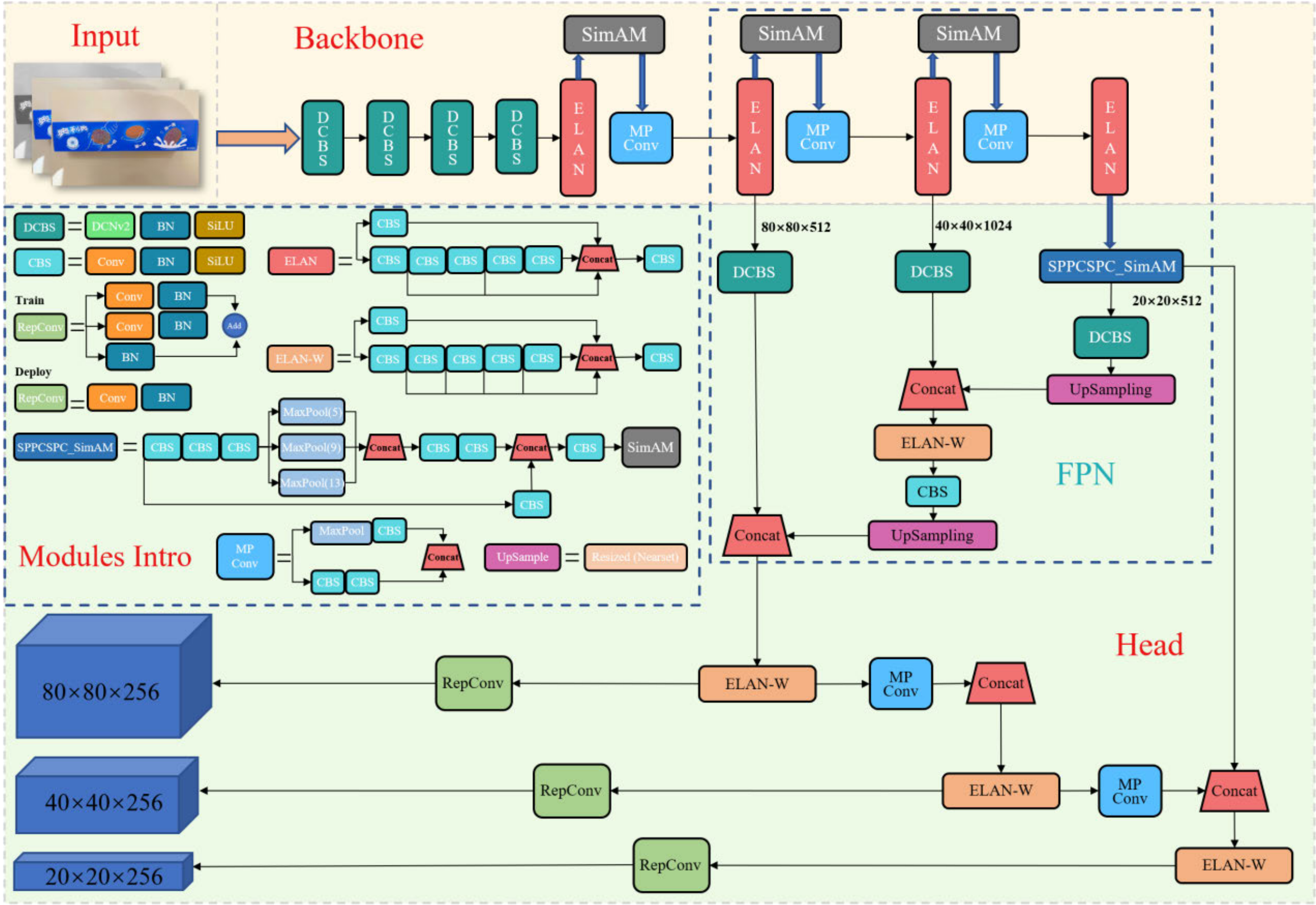

Input
Backbone
DCBS
ELAN
SimAM
MP Conv
80×80×512
40×40×1024
20×20×512
SPPCSPC_SimAM
UpSampling
Concat
ELAN-W
CBS
FPN
DCBS = DCNv2 BN SiLU
CBS = Conv BN SiLU
Train
RepConv = Conv BN Conv BN BN Add
Deploy
RepConv = Conv BN
ELAN = CBS CBS CBS CBS CBS CBS Concat CBS
ELAN-W = CBS CBS CBS CBS CBS CBS Concat CBS
SPPCSPC_SimAM = CBS CBS CBS MaxPool(5) MaxPool(9) MaxPool(13) Concat CBS CBS Concat CBS SimAM CBS
MP Conv = MaxPool CBS CBS CBS Concat
UpSample = Resized (Nearset)
Modules Intro
Head
RepConv
80×80×256
40×40×256
20×20×256

FIGURE 2. BP-YOLO network framework diagram. Modules Intro explains the details of each module in the network, and the optimized FPN structure is also reflected in the network diagram.

Finally, to avoid the serious imbalance of positive and negative sample ratio caused by foreground and background class imbalance in one-stage object detection model training process, cross entropy loss function in model is replaced by Focal Loss loss function. Focal Loss loss function based on binary cross entropy is a cross entropy loss function with dynamic scaling change. By adding dynamic scaling positive and negative sample weight factors and hard-easy sample modulation coefficients, it reduces sample weight of easy region in training process and refocuses on hard-to-distinguish sample weight. In optimizing direction of positive and negative sample imbalance, it improves model's overall detection performance. BP-YOLO network framework and above-mentioned modules are shown in Figure 2.

1) SIMAM ATTENTION MECHANISM

Inspired by the visual neuroscience theory, Yang et al. [34] proposed SimAM attention mechanism module (as shown in Figure 3), which can effectively enhance the expression of important feature information in three-dimensional space, and flexibly learn the important feature information in different spaces and channels. Compared with the existing channel attention mechanism and spatial attention mechanism that focus on two-dimensional weights, it has more advantages. In visual neuroscience, neurons that contain important information are in a unique active discharge mode, and such active neurons are given higher priority to suppress the activity of surrounding inactive neurons, showing a significant spatial inhibition effect. When applied to tasks, based on the linear separability between the target neuron and other neurons, the authors proposed an energy function to find the target active discharge neuron, as shown below:

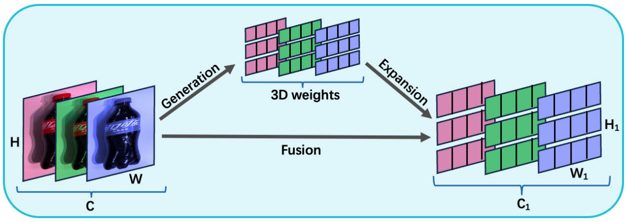


FIGURE 3. Structure of the SimAM 3D attention mechanism.

$$e_t\left(w_t, b_t, y, x_i\right) = \left(y_t - \hat{t}\right)^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}\left(y_o - \hat{x}_i\right)^2 \quad (1)$$

In the above equation, t and x_i are the active target discharge neurons and the surrounding non-target in the channel i of the input feature respectively. After linear transformation, $\hat{t} = w_t t + b_t$, $\hat{x}_i = w_t x_i + b_t$. Where w_t, b_t, y_o, y_t and M are: the weights, biases, numbers of active target discharge neurons and non-target neurons in the single channel, and the total number of neurons in the single channel, respectively. This equotion is equivalent to finding the linear separability between the target neuron t and other neurons in the single channel, and the energy value e_t is proportional to the importance of the target neuron $\hat{t}$. The energy function after adding the regularization parameter λ to the (1) is expressed by (2):

$$e_t\left(w_t, b_t, \mathbf{y}, x_i\right) = \frac{1}{M-1} \sum_{i=1}^{M-1} \left(-1 - \left(w_t x_i + b_t\right)\right)^2 + \left(1 - \left(w_t t + b_t\right)\right)^2 + \lambda w_t^2 \tag{2}$$

where the analytic solution for weight w_t and bias b_t is calculated by (3)-(4):

$$w_t = -\frac{2\left(t - \mu_t\right)}{\left(t - \mu_t\right)^2 + 2\sigma_t^2 + 2\lambda} \tag{3}$$

$$b_t = -\frac{1}{2}\left(t + \mu_t\right) w_t \tag{4}$$

Combining statistics, $\hat{\mu} = 1/(M-1)\sum_{i=1}^{M-1} x_i$ and $\hat{\sigma}^2 = 1/(M-1)\sum_{i}^{M-1}\left(x_i - \mu_t\right)^2$ can correspond to the distribution of mean and variance of non-target neurons in a single channel. Assuming that all pixels in a single channel follow the same distribution, then all neurons in a single channel follow this distribution, and the mean and variance of all neurons are calculated [1]. To avoid repeated calculations, the energy function can be expressed by (5):

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{\left(t - \hat{\mu}\right)^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{5}$$

The current energy function shows that the larger the energy, the greater the difference between the target neuron and the surrounding neurons, and the stronger its importance. The target neuron is distinguished from the surrounding neurons by means of the strength of the energy; and the relationship between the importance and the energy of each neuron is $t = 1/e_t^*$.

The introduction of the SimAM attention mechanism can effectively assist BP-YOLO in paying more attention to the product objects' three-dimensional spatial feature information in the training downsampling process, and reduce the loss of feature information in the transmission process, suppressing the noise interference caused by the complex environment around the product objects. While maintaining the original magnitude of the model, it improves the model's feature information perception and learning ability.

2) DEFORMABLE CONVNETS V2

Zhu et al. [2] proposed Deformable ConvNets v2 (DCNv2) based on the traditional deformable convolution, that introduces a modulation module to the convolution kernel, to achieve a reasonable control of the sampling range of the deformable convolution kernel. After offset modulation, the sampling points stay within an ideal feature range. The modulation module contains learnable offset parameters Δp (offset) and variation amplitude Δm. In addition, the stacking of multiple deformable convolution modules with modulation modules can enhance the model's robustness and feature information perception intensity for sparse space, and also play a role in correcting and refining the accuracy of offset and resisting spatial changes. This plays a key role in the field of dynamic video super-resolution.

Assuming that the size of the deformable convolution kernel is n×n, p_k and w_k are the position and the corresponding convolution kernel parameter of the k-th (k<n)element in the single convolution kernel, respectively. p_0 is the position on the output feature map. After the input feature map X passes through this module, the obtained feature module Y can be expressed by the (6):

$$Y\left(p_0\right) = \sum\nolimits_{k=1}^{K} X\left(p_0 + p_k + \Delta p_k\right) \cdot \Delta m_k \tag{6}$$

where Δp_k and Δm_k represent the offset parameters and variation amplitude that are learned by the network in this module, respectively. Δp_k is an unrestricted variable, and Δm_k has a variable range of [1, 0] after learning. In the model implementation, a sigmoid non-zero mean function is used to express the equation, as shown in Figure 4. The flexible-and-variable convolution kernel adjusts the receptive field finely through a modulable scalar, so that the new sampling position can better match the target interest area.

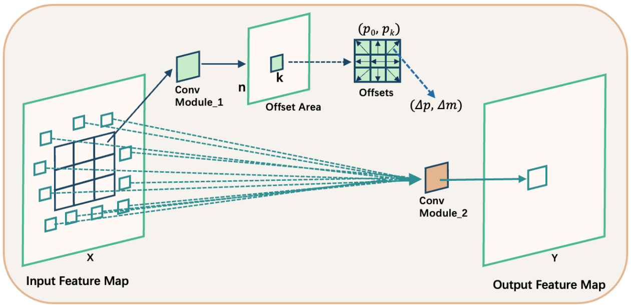


FIGURE 4. Deformable Convolution when the kernel size of the Conv Module is 3 × 3.

To address the deformation and partial occlusion of the target products packaging in the detection task, DCNv2 is used to replace the rigid convolution after the BP-YOLO input layer, and the offset can be added to the sampling points, so that the feature sampling range aligns with the object products feature area, achieving object reconstruction. In the FPN (Feature Pyramid Network) cascade, DCNv2 enlarges the semantic information receptive field of the low-resolution features in the cascade transmission, and improves the model's attention to the sparse space feature information. Therefore, the use of DCNv2 enhances the model's accurate sampling of the products features with packaging deformation and partial occlusion, and effectively reduces the probability of mis-detection and misssed-detection by the model.

3) FOCAL LOSS

The Focal Loss loss [3] function based on cross-entropy is applied in the one-stage object detection algorithm, which can be used as a solution to the unoptimized distribution of positive and negative samples. The core idea is to focus the gradient parameters updated in the model's backpropagation on the hard-to-detect samples, that is, the parameter optimization changes towards the hard-to-detect training samples. The loss function is defined as (7):

$$FL(p_t) = -\alpha_t(\left(1-p_t\right)^{\gamma} \log p_t) \tag{7}$$

where the $p_t = p$ if $y = 1$, otherwise $p_t = 1-p$. t is the sample index. α_t is the class balance weight factor, which suppresses the imbalance of positive and negative samples and controls the ratio of positive and negative sample losses. γ is the modulation coefficient, which reduces the loss of easy-to-classify samples and focuses on error-prone and hard-to-classify samples when the value is greater than 0. $(1-p_t)$ is the modulation factor, p is the overlap rate of the model's predicted region and the target region, then p_t can reflect the difficulty level of the model's target detection for a sample in the detection task. The larger p_t is, the higher the sample classification confidence is; on the contrary, the lower the sample classification confidence is, the harder the sample is to distinguish.

Therefore, in the training process of Focal Loss one-stage detection model, by increasing the weights of hard-to-detect samples and inaccurate samples in the loss function, the loss function tends to hard-to-detect samples and accurate samples, thereby improving the model's overall performance for product detection.

C. SHOPPING BEHAVIORS RECOGNITION

1) BLAZEPOSE

Traditional pose estimation methods generate heatmaps for the detected keypoints in the image, which can finely track the coordinates and offsets of each human joint [4], [5]. However, due to the high degree of freedom and realistic occlusion of human behavior in practical detection tasks, using heatmaps in the real-time inference stage reduces the model's scalability. BlazePose proposes a novel topology structure, which combines heatmaps, offsets, and regression methods, and stacks a mini encoder-decoder network based on heatmaps and a subsequent regression encoder network [6]. The heatmaps are only used in the training stage. In the model training stage, the input image passes through five convolutional layers, and the last four layers output heatmaps, offset feature information maps, and keypoints, respectively. The use of heatmaps in the training stage can effectively control the embedding of lightweight frameworks, improve the model's inference efficiency and high scalability, and calculate the model training loss with offsets. In the inference stage, the heatmap generation module is removed in the model's forward propagation process, and all joint coordinates are regressed directly through the encoder-decoder network architecture based on the stacked hourglass structure, thereby improving the accuracy of all joint predictions. At the same time, the gradients from the regression encoder network do not affect the gradient calculation in the heatmap training process. BlazePose can run at about 30fps and track 33 key skeletal points on the human body in real time, as shown in Figure 5.

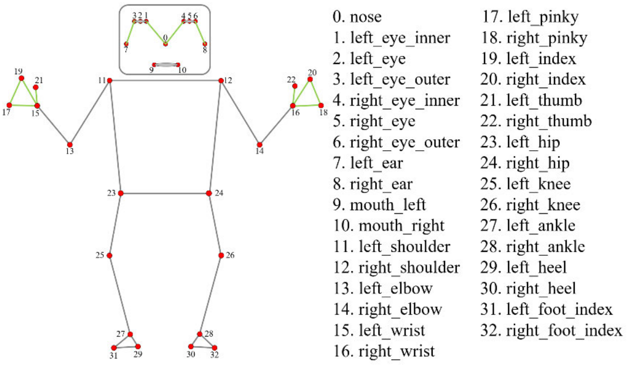


FIGURE 5. The 33 key skeletal points tracked by BlazePose.

The angles formed by the human joints are the key to determine the behavior in the process of human action recognition by visual pose algorithms. This study uses the BlazePose algorithm to track the wrist, elbow, and shoulder joints of consumers in videos, and the algorithm can return the pixel coordinate positions of each joint in each frame of the image. Based on the pixel positions, we can calculate the relative angles of each joint in the video, and define the corresponding behavior categories for different angle ranges between joints, thus achieving the recognition of human behavior. The angle calculation uses the method of inclination calculation, taking the human shoulder, elbow, and wrist joints as the fulcrum, and taking the angle between the upper arm and the lower arm as an example, as shown in Figure 6.

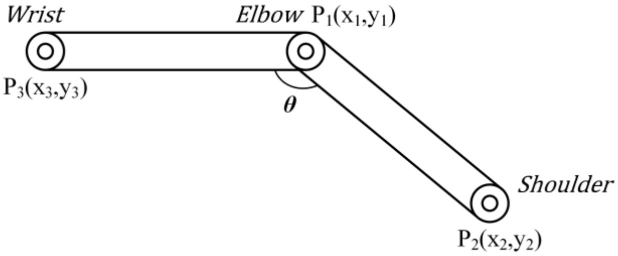


FIGURE 6. The shoulder, elbow, and wrist joints of one side of the human body, where $P_1(x_1, y_1)$ is the coordinate point of the elbow joint, $P_2(x_2, y_2)$ is the coordinate point of the shoulder joint, $P_3(x_3,y_3)$ is the coordinate point of the wrist joint, and the angle examples between the joints are shown.

As shown in Figure 6, P1 is set as the reference point, and P2 and P3 are the relative change points. Clearly, any change in P1, P2, or P3 can lead to the angle θ. In a specific scenario,

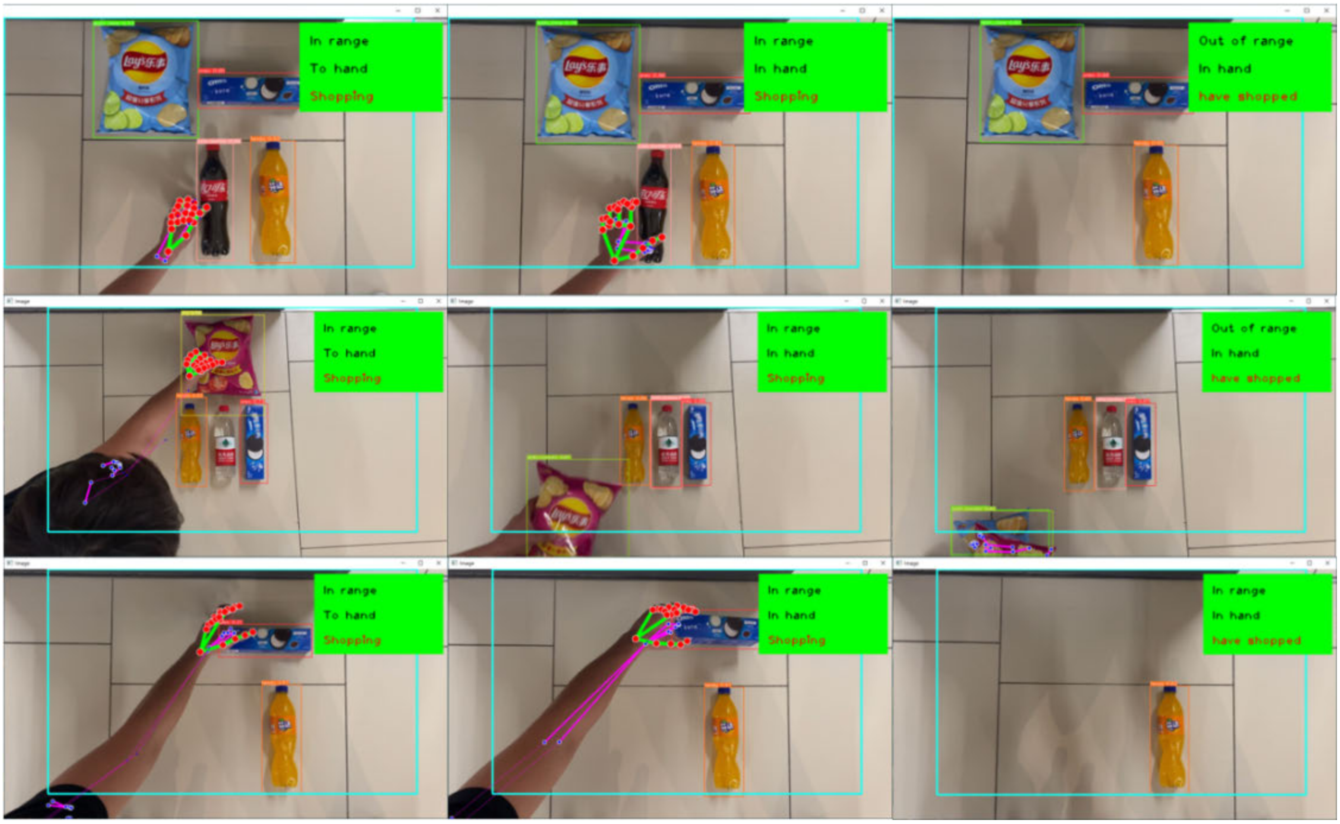


FIGURE 7. **Simulated shopping process. The blue box on the ground serves as the shopping area, and the green module is the shopping information display area.**

the current angle calculation can determine the behavior represented by the hand movement. For convenience in the test, the vector method is used to find the interior angle of the three points, and the angle between the two vectors $\overrightarrow{P_{12}}$ and $\overrightarrow{P_{13}}$ is shown in (8).

$$\theta = \arccos\left(\frac{\overrightarrow{P_{12}}.\overrightarrow{P_{13}}}{|\overrightarrow{P_{12}}||\overrightarrow{P_{13}}|}\right) \tag{8}$$

Substitute the coordinates of the points into the equation, we can calculate it as the (9).

$$\theta = arccos\frac{y_1^2 - y_1 \cdot y_2}{y_1\sqrt{(x_2 - x_1)^2 + (v_2 - v_1)^2}} \tag{9}$$

2) PRODUCTS AND SHOPPING BEHAVIORS RECOGNITION MODEL

This study proposes the BP-YOLO model by integrating the improved YOLOv7 product detection and the BlazePose pose estimation models. By simulating the process of consumers using smart unmanned cabinets to shop products in real life, BP-YOLO dynamically detects the products categories, and recognizes the consumer's behavior and shopping status of picking up products, thereby judging whether the consumer has successfully purchased a certain product.

The BP-YOLO algorithm uses the following logic to judge the consumer's purchase behavior: The algorithm splits the real-time transmitted video stream into image frames and uses OpenCV to segment an area according to the image ratio as a simulated shopping platform for the products placed in the simulated shopping cabinet. It also sets a green border as the shopping status display area, which includes Out of range (the consumer's hand does not enter the shopping area), In range (the consumer's hand enters the shopping area), To hand (the consumer's reaching state), In hand (the consumer's retracting state with a product), Shopping (shopping state) and Have shopped (successful purchase). When the video detects that the consumer's hand enters the image, the behavior detection algorithm will continuously return the pixel coordinate information of the consumer's finger, palm, arm and other joints, and the shopping status is displayed as: In range, To hand and shopping; The target detection algorithm will return the pixel coordinate information of the product category and product prediction box detected in each frame of the image. If the angle between the consumer's lower arm and upper arm is greater than 90 degrees, and the hand position enters the shopping area, it is judged that the consumer is reaching for a product; When the pixel position area composed of the consumer's palm or finger overlaps with the center coordinate of a product, it starts timing. if it lasts for more than 2s, it judges that the consumer is taking a certain product, and display the information as: In range, In hand and Shopping; If the product taking time is less than 2s, it re-judge whether the consumer's hand touches a product. When the model determines that the consumer has taken a certain product, if the angle between the consumer's lower arm and upper arm is less than 120 degrees, it belongs to a retracting state. At this point, if both the center position coordinates of a product and a hand leave the area of the simulated shopping platform, it judges that the consumer has purchased this product, and returns its category. The information display area is: Out of range, In hand and Have shopped. The real-time detection speed of this integrated model reaches about 36 frames per second, as shown in Figure 7.

IV. EXPERIMENT

A. DATASET

Our proposed method and model are trained, evaluated on our custom dataset. We collect 162 images of 15 types of high-selling products in supermarkets using artificial multi-angle shooting. Some products are captured in a simulated shopping state with hand-induced occlusion to mimic a realistic scenario. We enrich our dataset by applying batch grayscale changes, adding various noises, adjusting brightness, saturation and contrast, etc. to the images, resulting in 1620 augmented image data. For the product detection task, the input layer process of BP-YOLO can randomly crop, stitch and reorganize the images. We use Labelimg to annotate the object anchor boxes for the original 162 images (see Figure 8) according to the COCO dataset format, and generate JSON files containing product positions and categories accordingly. We reuse these annotation files for the remaining 1620 images as their relative positions and categories are unchanged. We split our dataset into training and validation sets with an 8:2 ratio.

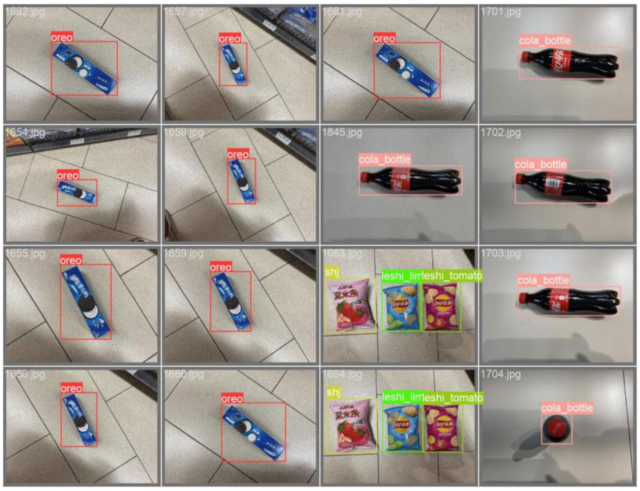


FIGURE 8. Product dataset examples. It contains five products, which are photographed from different angles and combinations, and annotated with category labels.

B. EXPERIMENT SETUP

We conducted our experiments on a PC workstation equipped with 256GB of memory and two NVDIA GeForce RTX4000 GPUs, each with 8GB of memory and 14 TFLOPS of processing power. We used Pytorch to implement and evaluate our project model. The model training parameters were as follows: batch size = 16, epochs = 300, initial learning rate = 0.001, and learning rate scheduler = LambdaLR. Moreover, we applied transfer learning to enhance the efficiency and accuracy of YOLOv7. We employed the Adam optimizer to update and optimize the model weights based on the pre-trained model.

C. RESULTS AND ANALYSIS OF PRODUCT DETECTION MODEL

For the evaluation the performance of our product detection model BP-YOLO, we used three metrics: single-class average precision *AP* (Average Precision), mean average precision *mAP* (mean-Average Precision), and average inference speed of the model for image detection. *AP* measures the area under the precision-recall curve of the model, and is calculated as follows.

$$AP = \int_0^1 P(r)\,dr \tag{10}$$

where $P(r)$ is the precision at a given recall r, the functions are as follows.

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$r = \frac{TP}{TP + FN} \tag{12}$$

where *TP* is the number of true positives and *FP* is the number of false positive.

$mAP@[0.5 : 0.95]$ is used to measure the model's performance in product category classification and localization, which is calculated IoU by with a step size of 0.05 in the range of 0.5 to 0.95. IoU [7] is the intersection over union of the model's predicted box and the ground truth box, which is used to determine the accuracy of the model's bounding box for object detection. The formulas are as follows.

$$mAP = \frac{1}{m}\sum\nolimits_{i=1}^{m} AP_i \tag{13}$$

$$IoU = \frac{B \cap G}{B \cup G} \tag{14}$$

where m is the number of product categories. The larger the IoU value, the higher the overlap between the model prediction box and the ground truth box, and the more accurate the model's prediction box compared to the ground truth box.

To evaluate the feasibility, authenticity and robustness of our model, we trained and tested the classic two-stage, one-stage models and BP-YOLO with different Backbones on our proposed custom dataset, under the same settings of epochs, batch size, optimizer and other hyperparameters. Table 1 shows that all models have slight differences in mAP@.5 and mAP@[.5, .95], highlighting the impact of IoU on detection performance and the necessity of considering this parameter range in our application context. Compared with one-stage

TABLE 1. A comparison of seven object detection models for product detect.

Model	Backbone	mAP@.5 (IoU=0.5)	mAP@[.5, .95] (IoU=[0.5, 0.95])	Average Inference Time (ms / image)
Fast R-CNN	VGG-16	85.35	82.20	158.70
Faster R-CNN	Resnet50	93.64	85.42	120.54
SSD	VGG-16	86.59	74.35	64.17
YOLOv3	DarkNet53	88.81	80.46	72.38
YOLOv4	DarkNet53	90.93	78.91	75.35
YOLOv7	CSPDarknet	94.61	88.39	19.21
BP-YOLO	**FIGURE 2**	**98.59**	**96.17**	**26.89**

models, Fast R-CNN and Faster R-CNN achieve better detection results in mAP, but their average inference time increases by at least three times, which is not suitable for real-time detection tasks. In the mAP comparison of one-stage models, YOLOv7 outperforms SSD, YOLOv3 and YOLOv4 models in both results and inference efficiency. BP-YOLO, based on optimized YOLOv7, introduces SimAM and DCNv2 modules, which increase model parameters and computation, but achieve the best results in mAP and do not affect detection efficiency significantly.

These results indicate that YOLOv7 is more suitable for product detection than other models, and demonstrate the superiority of our proposed BP-YOLO as the final product detection model.

In the real application scenarios, the products in the UVM cabinet are not single entities, but products of different shapes and categories placed together. Therefore, AP is used as the evaluation metric, and the comparison of the detection accuracy results of YOLOv7 and BP-YOLO for each category is shown in Figure 9.

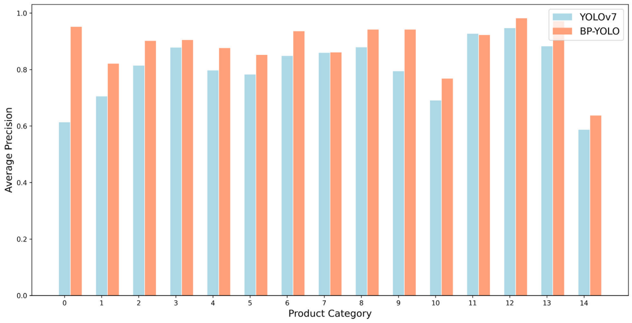


FIGURE 9. Comparison of the average detection precision of YOLOv7 and BP-YOLO for 15 categories of products.

As shown in Figure 10, when the model is trained to the 300th epoch under the above mAP and IoU settings, the mAP@[0.5:0.95] of BP-YOLO reaches 96.17%, which is nearly 8% higher than that of YOLOv7.

In the above experiment, the dataset included product images that were affected by hand occlusion, packaging wrinkles, overexposure and underexposure. The baseline model achieved a mAP@[0.5:0.95] of 88.59% on this dataset. This means that the native model was likely to have missed or mis-detection in the dynamic detection process of product images in similar complex environments. To address this issue, BP-YOLO added 3D parameter-free attention mechanism SimAM and deformable convolution DCNv2, which effectively reduced the information loss in the feature transmission process, enhanced the model's feature perception of the target products in the image, and improved the average detection accuracy of products under IoU from 0.5 to 0.95.

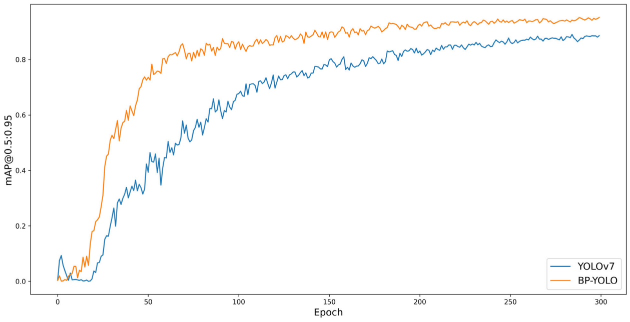


FIGURE 10. Comparison of mAP between YOLOv7 before and after improvement under IoU from 0.5 to 0.95aption.

D. RESULTS AND ANALYSIS OF SHOPPING RECOGNITION

We propose to utilize BP-YOLO to achieve parallel recognition of consumer's product category and behavior during the shopping process. Following Zhang et al., who proposed the success rate as a performance metric to evaluate the model's ability to simultaneously predict the correct category and instance number of products in their experiment, we use the same metric to further validate the accuracy of our integrated model in product detection and shopping behavior recognition in dynamic shopping scenarios. The success rate is defined as the ratio between the number of times that the integrated model correctly judges whether the consumer has purchased a certain product and the total number of simulated shopping trials in the consumer's shopping process. In this case, the result of the integrated model is considered successful only if it correctly predicts the product category and correctly judges the consumer's purchase behavior. We simulate three consumers to purchase any product from 15 categories and record the real-time recognition results of the model under three different lighting and noise environments: high, medium, and low. The number of frames captured in each environment is around 30, and each consumer simulates picking up products 50 times. In the experiment, we compare and analyze the success rate of each consumer in different environment settings when picking up products, as shown in Table 2 below.

TABLE 2. Shopping recognition success rate under different experimental environments.

Light condition Experiment situation	High light	Normal light	Low light	Average success rate
Customer (High noise)	90%	94%	92%	92%
Customer (Normal noise)	98%	100%	96%	98%
Customer (Low noise)	94%	96%	94%	94.7%

BP-YOLO achieved an average success rate of 92 %, 98%, and 94.7% in correctly identifying the types and behaviors of consumer purchases under three different lighting and noise intensities, respectively, according to the experimental records. The results indicate that the model BP-YOLO can accurately identify the categories and shopping behaviors of dynamically picked products by consumers in different lighting environments and situations where the consumer's hand partially occludes or deforms the product packaging. This also further tests the robustness and stability of the model, demonstrating its potential for deployment in realistic consumption scenarios.

V. DISCUSS

This section discusses the limitations of BP-YOLO and provides some suggestions for future work to meet a higher efficiency and accuracy requirements of intelligent UVMs in commercial deployment.

1) BP-YOLO employs DCNv2 that leverages deformable convolution kernels, incorporates more deformable convolution layers and offset learning modulation mechanisms, confers the model with the capability of flexible feature sampling, and substantially enhances the model's performance for sparse spatial feature learning. However, the model needs not only to learn the weights of the sampling points, but also to learn the offset and offset amplitude parameters, which incur increased costs of training time, learning parameters, computation, and memory overhead. Therefore, inspired by the model optimization scheme of dynamic weight slicing proposed by Li et al [42], it mainly slices a part of network parameters dynamically according to the difficulty level of the input, thus achieving different model capacities. Without losing too much accuracy and keeping the parameters statically and continuously stored in the hardware, it can avoid the extra overhead brought by sparse computation and reduce the model computation. In the future, we will continue to further study and implement this operation into our model.
2) BP-YOLO enables consumers to take out multiple products at a time based on dynamic vision. However, the current product training set is small and the product categories are not diverse enough, which may limit the robustness and generalization of the model. Therefore, it is necessary to further enrich and expand the data set work to cover more product types and scenarios.
3) Inspired by the ContrastZSD model proposed by Yan et al. [43], which introduces contrastive learning mechanism into the zero-shot object detection framework to learn more knowledge about unseen categories and optimizes the visual data structure. In the future, our proposed model will adopt this model to use a large amount of known product information, eliminating the need for labeling new products from unknown and updated brands. We use the aligned features to recognize product and behavior recognition, making the model adapt to the diversity and novelty of products, and improving the flexibility and discrimination of recognition.
4) We used a normal camera to record the experimental picking process in the shopping behavior recognition experiment, and the recording video suffered from severe motion blur effect problem. Since the BlazePose model is sensitive to body part keypoint detection and skeleton generation, a binocular camera, fisheye or other high-speed camera can be used to provide clearer series of frames for action and content analysis, which may improve the accuracy of determining the consumer's shopping behavior. Moreover, BP-YOLO can eliminate the use of RFID and other infrared physical sensors, but impose higher requirements for hardware such as processors, memory, and graphics cards for vision-based behavior recognition models. Therefore, it is important to consider the trade-off between hardware cost and performance when deploying BP-YOLO in commercial settings.
5) In order to improve and balance the model's learning, detection accuracy and inference efficiency, it is still necessary to pay attention to the improvement of BP-YOLO model network lightweighting, structure design adjustment and related hardware enhancement in the future, so as to further boost its speed, performance and stability in actual application scenarios.
6) The intelligent UVM proposed in this paper has the advantages of scalability and elasticity in future commercial deployment, which are mainly reflected in the following software and hardware aspects. In terms of software, we can transform the algorithm model into ONNX model representation and perform inference on the product category and shopping behaviors on a professional inference engine framework, such as jetson

nano, which avoids cloud computing and servers, and achieves fast computation on the local terminal, with a speed increase of 2-4 times. In terms of hardware, we can adopt standardized modular design for UVM in the future, which can be flexibly combined and split according to different scenarios and needs, and realize the rapid deployment and movement of the vending machine. At the same time, we can also dynamically adjust the number and location of the UVM according to the real-time traffic and sales, and balance the utilization and cost of the vending machine.

VI. CONCLUSION

This paper aims to address the risk of mis-detection and missed-detection of products in complex shopping environments of UVM, and to provide a more vision-based intelligent UVM shopping settlement method. Based on the existing object detection and tracking techniques, this paper proposed a more robust and innovative BP-YOLO product and shopping behavior recognition model. This paper validates the effectiveness and reliability of the algorithm through extensive experiments, and its main contributions and innovations are as follows:

(1) In the product detection task, the BP-YOLO model's mAP@[0.5:0.95] is 96.17%, which is about 8% higher than the original YOLOv7, and its excellent product detection accuracy can meet the dynamic product detection task in various scenarios.

(2) In the shopping recognition experiment, the BP-YOLO model performs real-time detection and tracking of the products and consumers' shopping behavior inside the UVM cabinet, and determines whether the consumer has purchased certain products. This paper simulates various realistic shopping scenarios, and conducts comprehensive experiments in three different light and noise environments: high, medium, and low. The success rate is used as the evaluation metric, and the average recognition success rate can reach 92%, 98%, and 94.7%, respectively, and the detection speed is maintained at about 40fps, meeting the requirements of implementation and stability.

(3) The intelligent UVM proposed in this paper has strong practicality and commercial value, and can provide consumers with a convenient, fast, and secure shopping experience, save manpower and material costs for vending machine operators, and contribute to the popularization and development of unmanned vending machines. The model proposed in this paper can also be extended to other similar scenarios, such as unmanned supermarkets, unmanned hotels, etc., and has a broad application prospect.

REFERENCES

[1] S. Chen, "Design of commodity and shopping behavior recognition algorithm based on retail model of unmanned vending machine," M.S. thesis, Jinan Univ., 2020, doi: 10.27167/d.cnki.gjinu.2020.001918.

[2] J. Xu, Z. Hu, Z. Zou, J. Zou, X. Hu, L. Liu, and L. Zheng, "Design of smart unstaffed retail shop based on IoT and artificial intelligence," *IEEE Access*, vol. 8, pp. 147728–147737, 2020, doi: 10.1109/ACCESS.2020.3014047.

[3] F. Shang, P. Yang, J. Xiong, Y. Feng, and X. Li, "Tamera: Contactless commodity tracking, material and shopping behavior recognition using COTS RFIDs," *ACM Trans. Sensor Netw.*, vol. 19, no. 2, pp. 1–24, Feb. 2023, doi: 10.1145/3563777.

[4] K. Xia, H. Fan, J. Huang, H. Wang, J. Ren, Q. Jian, and D. Wei, "An intelligent self-service vending system for smart retail," *Sensors*, vol. 21, no. 10, p. 3560, May 2021, doi: 10.3390/s21103560.

[5] B. Wong, "Design and implementation of intelligent retail cabinet software system based on gravity sensing and target detection," Southwest Univ., Tech. Rep., 2021, doi: 10.27684/d.cnki.gxndx.2021.002250.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[7] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[10] G. Ye, J. Qu, J. Tao, W. Dai, Y. Mao, and Q. Jin, "Autonomous surface crack identification of concrete structures based on the YOLOv7 algorithm," *J. Building Eng.*, vol. 73, Aug. 2023, Art. no. 106688, doi: 10.1016/j.jobe.2023.106688.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[13] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[14] M. Yin, X. Jia, X. Zhang, J. Feng, and X. Fan, "Human detection of damage behavior for vending cabinets based on improved YOLOv4-tiny," *Comput. Eng. Appl.*, pp. 1–10.

[15] S.-J. Horng and P.-S. Huang, "Building unmanned store identification systems using YOLOv4 and Siamese network," *Appl. Sci.*, vol. 12, no. 8, p. 3826, Apr. 2022, doi: 10.3390/app12083826.

[16] L. Liu, J. Cui, Y. Huan, Z. Zou, X. Hu, and L. Zheng, "A design of smart unmanned vending machine for new retail based on binocular camera and machine vision," *IEEE Consum. Electron. Mag.*, vol. 11, no. 4, pp. 21–31, Jul. 2022, doi: 10.1109/MCE.2021.3060722.

[17] B. Zhu, G. Xiao, Y. Zhang, and H. Gao, "Multi-classification recognition and quantitative characterization of surface defects in belt grinding based on YOLOv7," *Measurement*, vol. 216, Jul. 2023, Art. no. 112937, doi: 10.1016/j.measurement.2023.112937.

[18] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," 2020, *arXiv:2006.10204*.

[19] Z. Ke, Z. Chen, H. Wang, and L. Yin, "A visual human-computer interaction system based on hybrid visual model," *Secur. Commun. Netw.*, vol. 2022, Jun. 2022, Art. no. e9562104, doi: 10.1155/2022/9562104.

[20] W. Liu, X. Liu, Y. Hu, J. Shi, X. Chen, J. Zhao, S. Wang, and Q. Hu, "Fall detection for shipboard seafarers based on optimized BlazePose and LSTM," *Sensors*, vol. 22, no. 14, p. 5449, Jul. 2022, doi: 10.3390/s22145449.

[21] Y.-P. Huang, S. Kshetrimayum, and C.-T. Chiang, "Object-based hybrid deep learning technique for recognition of sequential actions," *IEEE Access*, vol. 11, pp. 67385–67399, 2023, doi: 10.1109/ACCESS.2023.3291395.

[22] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[23] M. Klasson, C. Zhang, and H. Kjellström, "A hierarchical grocery store image dataset with visual and semantic labels," 2019, *arXiv:1901.00711*.

[24] H. Zhang, D. Li, Y. Ji, H. Zhou, W. Wu, and K. Liu, "Toward new retail: A benchmark dataset for smart unmanned vending machines," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7722–7731, Dec. 2020, doi: 10.1109/TII.2019.2954956.

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

[26] S. Chen, D. Liu, Y. Pu, and Y. Zhong, "Advances in deep learning-based image recognition of product packaging," *Image Vis. Comput.*, vol. 128, Dec. 2022, Art. no. 104571, doi: 10.1016/j.imavis.2022.104571.

[27] H. Sun, J. Zhang, and T. Akashi, "TemplateFree: Product detection on retail store shelves," *IEEJ Trans. Electr. Electron. Eng.*, vol. 15, no. 2, pp. 242–251, Feb. 2020, doi: 10.1002/tee.23051.

[28] K. S. Changan and P. G. Chilveri, "Stereo image feature matching using Harris corner detection algorithm," in *Proc. Int. Conf. Autom. Control Dynamic Optim. Techniques (ICACDOT)*, Sep. 2016, pp. 691–694, doi: 10.1109/ICACDOT.2016.7877675.

[29] O. Haggui, C. Tadonki, L. Lacassagne, F. Sayadi, and B. Ouni, "Harris corner detection on a NUMA manycore," *Future Gener. Comput. Syst.*, vol. 88, pp. 442–452, Nov. 2018, doi: 10.1016/j.future.2018.01.048.

[30] D. Li, H. Zhou, G. Li, B. Yang, F. Gao, and H. Zhang, "Drt-Net: An improved RetinaNet for detecting beverages in unmanned vending machines," in *Proc. IEEE Int. Symp. Product Compliance Eng.-Asia (ISPCE-CN)*, Nov. 2020, pp. 1–6, doi: 10.1109/ISPCE-CN51288.2020.9321854.

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.

[32] M. Hu, Y. Li, L. Fang, and S. Wang, "A^2-FPN: Attention aggregation based feature pyramid network for instance segmentation," May 2021, *arXiv:2105.03186*.

[33] H. Park and J. Paik, "Pyramid attention upsampling module for object detection," *IEEE Access*, vol. 10, pp. 38742–38749, 2022, doi: 10.1109/ACCESS.2022.3166928.

[34] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 11863–11874. [Online]. Available: https://proceedings.mlr.press/v139/yang21o.html

[35] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative decorrelation for clustering and classification," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2012, pp. 459–472, doi: 10.1007/978-3-642-33765-9_33.

[36] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," 2018, *arXiv:1811.11168*.

[37] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[38] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," 2019, *arXiv:1903.06593*.

[39] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*.

[40] S.-T. Kim and H. J. Lee, "Lightweight stacked hourglass network for human pose estimation," *Appl. Sci.*, vol. 10, no. 18, p. 6497, Sep. 2020, doi: 10.3390/app10186497.

[41] S. Wu, J. Yang, X. Wang, and X. Li, "IoU-balanced loss functions for single-stage object detection," *Pattern Recognit. Lett.*, vol. 156, pp. 96–103, Apr. 2022, doi: 10.1016/j.patrec.2022.01.021.

[42] C. Li, G. Wang, B. Wang, X. Liang, Z. Li, and X. Chang, "DS-Net++: Dynamic weight slicing for efficient inference in CNNs and vision transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4430–4446, Apr. 2023, doi: 10.1109/TPAMI.2022.3194044.

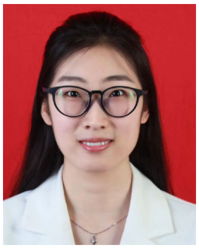
[43] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 4, 2022, doi: 10.1109/TPAMI.2021.3140070.

JINGXIANG LI received the B.S. degree in computer science and information. He is currently pursuing the M.S. degree with the Xi'an University of Science and Technology. He has published one article and one software copyright in related fields. He has also won provincial and national awards in mathematical modeling contests. His research interests include deep learning and computer vision.

FUQUAN TANG received the Ph.D. degree in engineering from the Xi'an University of Science and Technology, Xi'an, China, in 2009. He is currently a Professor and a Ph.D. Supervisor with the School of Geomatics, Xi'an University of Science and Technology, where he is also the Leader of the Doctoral Program in Surveying and Mapping Science and Technology. He has presided over six provincial and ministerial level scientific research projects, including the National Natural Science Foundation of China, completed more than 30 coal mine surveying and mining subsidence engineering projects, and won four provincial and ministerial level scientific and technological progress awards. He has published more than 60 academic articles, five monographs and textbooks, and obtained 11 invention and utility model patents and nine software copyrights. He is a member of the Mine Surveying Professional Committee and the Education Committee of the Chinese Society for Geodesy, Photogrammetry and Cartography, the Director of the Underground Engineering and Mine Surveying Special Committee of Shaanxi Province, a member of the Mine Surveying Professional Committee of the China Coal Society, and a member of the Engineering Professional Accreditation Expert Group of the Ministry of Education.

CHAO ZHU is currently pursuing the Ph.D. degree with the Xi'an University of Science and Technology, China. He has published four articles in peer-reviewed journals. His research interests include airborne LiDAR surveying, deep learning, and computer vision.

SHIWEI HE received the B.Eng. degree in safety engineering from the China University of Mining and Technology, China, in 2021. He is currently pursuing the Master of Science degree in information systems with New York University, USA. He has applied his skills to various projects, such as unsafe behavior recognition in coal mining environments using YOLOv5 algorithms, real-time image stitching with OpenCV using SIFT and SURF algorithms, and feature disentanglement using TL-GAN. He has published one article on TL-GAN in a peer-reviewed journal. His research aims to improve industrial safety protocols using machine learning techniques. His research interests include algorithms, machine learning, database systems, and computer vision.



SHUJIN ZHANG received the B.E. degree in computer science from Inner Mongolia Agricultural University, Hohhot, China, in 2018. She is currently pursuing the M.A. degree in agricultural engineering with Northwest A&F University, Yangling, Shaanxi, China. Her research interests include artificial intelligence, image processing, and intelligent livestock and poultry farming technology.

YU SU received the B.S. degree in computer science and technology. He is currently pursuing the M.S. degree. In 2023, he published one article related to the computer vision. His main research interests include UAV photogrammetry and computer image processing.