American Journal of Neural Networks and Applications
2026, Vol. 12, No. 1, pp. 25–33
https://doi.org/10.11648/j.ajnna.20261201.13

SciencePG
Science Publishing Group

Research Article

The Retail Product Recognition for Intelligent Unmanned Vending Machines Using YOLO and Enhanced Swin Transformer

Kwang-Min Kim, Kwang-Uk Han, Song-Jun Yun, Hakho Hong*

Institute of Mathematics, State Academy of Sciences, Pyongyang, Democratic People's Republic of Korea

Abstract

Recently, with the rapid development of deep learning in the retail industry, intelligent unmanned vending machines (UVMs) have become a dominant trend in retail product applications. The most advanced retail product recognition systems for intelligent s are based on object detection models such as Fast R-CNN or YOLO, and image classification models using CNNs like ResNet. However, the complexity of practical environments makes intelligent UVMs face challenges in product recognition performance and many studies are exploring ways to improve this performance. Recently, Vision based Transformers (ViTs) outperform the traditional convolutional neural networks (CNNs) like ResNet, particularly Swin Transformer has gained attention as it is well suited on small dataset and shows promising results compared to the other ViTs. In this paper, we propose two-stage pipeline consisting of product detection and recognition models. The proposed approach utilizes YOLO11 for product detection and an enhanced Swin Transformer for product recognition. Enhanced Swin Transformer is consisted of two main modules: multi-level feature fusion module and a global multi-scale attention module to improve product recognition performance of original Swin Transformer. This proposed approach combines both stages of product detection and recognition into a unified pipeline, leveraging the advantages of each method while avoiding the limitations imposed by single-stage detection models. We demonstrate the effectiveness of the proposed system through the several experiments. The proposed model improves True Acceptance Rate (TAR) by 1.2% when False Acceptance Rate (FAR) is equal to 1e-3. The proposed retail product recognition system using enhanced Swin Transformer demonstrates remarkable generalization capabilities and can handle boundless products without retraining the models.

Keywords

Intelligent Unmanned Vending Machine, Retail Product Recognition, Product Detection, Swin Transformer, Global Multi-Scale Attention, YOLO

1. Introduction

UVMs in the retail industry can be classified into two types: traditional UVMs and intelligent UVMs. Traditional UVMs based on automation technology deliver only beverages and they have limitations that a customer cannot touch the products due to being locked inside the machine and products cannot be returned when he does not want them. On the other

*Correspondence: Hakho Hong (hhhong@star-co.net.kp)

Received: 27 November 2025; **Accepted:** 4 January 2026; **Published:** 14 April 2026

hand, intelligent UVMs allow customers to open the door, inspect products by checking their expiration dates and reading product information, and place something back if they do not want (we refer to intelligent UVM as UVM in what follows) [10, 12, 15].

The advanced computer vision techniques used to make UVMs evolved from static vision to dynamic vision, but they still face challenges like product package deformation, noise and partial occlusion from the real-world scenarios. Several studies have been conducted to improve their performance.

The most advanced product recognition systems for UVMs are based on object detection models such as Fast R-CNN or YOLO, and image classification models using CNNs like ResNet [2, 4, 5, 11, 13]. [13] proposed a product detection based on Faster R-CNN object detection model with ResNet50 backbone. A two-stage object detection and recognition pipeline which composed of product detection based on Faster-RCNN and image encoder based on ResNet-18 for product recognition was proposed in [11]. In [2, 5] the traditional one-stage object detection models, YOLO series were deployed in product detection for UVMs. [4] proposed BP-YOLO based on enhanced YOLOv7 object detection and BlazePose pose estimation.

On the other hand, ViTs outperform CNNs like ResNet, particularly Swin Transformer has become a foundational architecture for vision tasks with its hierarchical structure and efficient self-attention mechanism and several studies using Swin Transformer have been conducted [6-9]. In [7], they proposed Swin Transformer and achieved the SOTA results on image classification, object detection and semantic segmentation tasks. Its performance surpassed the previous ViTs and CNNs by large margins. [9] presented a multi-purpose algorithm for simultaneous face recognition, facial expression recognition, age estimation, and face attribute estimation based on a single Swin Transformer and a Multi-Level Channel Attention (MLCA) module. The proposed method achieved the SOTA performances in facial expression recognition and age estimation respectively.

In this paper, we propose two-stage pipeline consisting of product detection and recognition models. The proposed approach utilizes YOLOv11 for product detection and an enhanced Swin Transformer for product recognition which consists of two main modules: Multi-Level Feature Fusion (MLFF) [10] module and a Global Multi-Scale Attention (GMSA) module. MLFF is used to combine different scale feature maps from each stage of Swin Transformer, while the final output is formed by summing up the global feature output and the output obtained through the GMSA from the MLFF combination.

The rest of this paper is organized as follows. In Section 2, we briefly summarize BP-YOLO and Swin Transformer. In Section 3, we propose an enhanced Swin Transformer with GMSA and develop a product recognition system for UVMs. Experiments in Section 4 demonstrate the effectiveness of the proposed methods. The main finding of this paper is concluded in Section 5.

2. Related Work

In [4], they proposed a BP-YOLO model that incorporates YOLOv7 [14] for product recognition and BlazePose [1] for shopping behaviors recognition by introducing the 3D attention mechanism SimAM and the deformable ConvNets v2 (DCNv2) (Figure 1). They added 3D SimAM attention mechanism module to the ELAN module in the Backbone layer, and SPPCSPC module in the Feature Pyramid Network (FPN) structure is replaced by SPPCSPC_SimAM module and introduced DCBS to make the model more attentive to sparse spatial information.

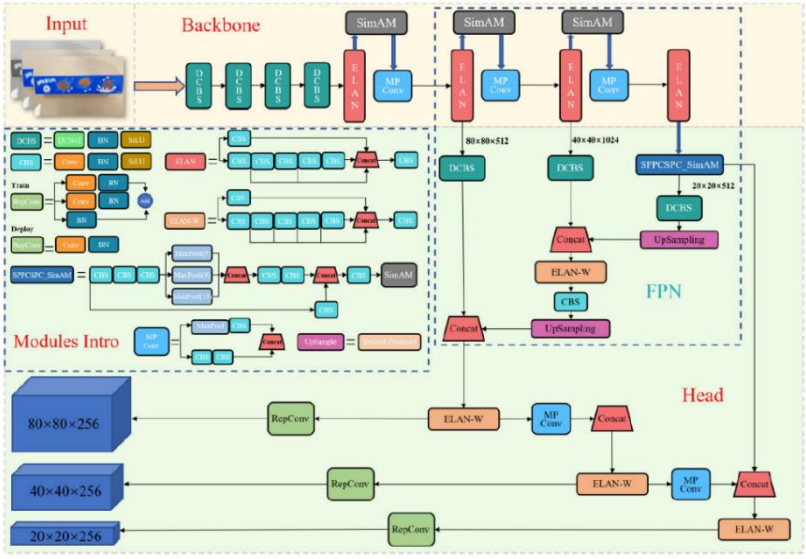

Figure 1. *BP-YOLO network diagram.*

However, the product recognition based on state-of-the-art object detection is able to classify only a certain number of products, and requires a lot of efforts to annotate images and should be retrained to handle new products. A practical approach should be able to handle new products without retraining the model from scratch.

To tackle the above issues, we consider product recognition by a two-stage pipeline consisting of product-agnostic detection and product recognition through a similarity search between feature embeddings for reference database and cropped images. It is crucial to develop a model that can extract discriminative feature embeddings for arbitrary products.

The Swin Transformer is a hierarchical Vision Transformer designed to improve efficiency, scalability, and performance for computer vision tasks such as image classification, object detection, and image segmentation. Its hierarchical approach allowed Swin Transformer to outperform the other ViTs and to address the computational inefficiency of standard ViTs.

Swin Transformer is built by replacing the standard multi-head self-attention (MSA) module in a Transformer block by the shifted window attention module composed of window based multi-head self-attention (W-MSA), where self-attention is computed within non-overlapping local windows to reduce complexity, and shifted window based multi-head self-attention (SW-MSA) that allows information exchange between neighboring windows. This enables Swin Transformer to capture both local and global dependencies while significantly reducing computational cost compared to standard ViTs using global MSA (Figure 2).

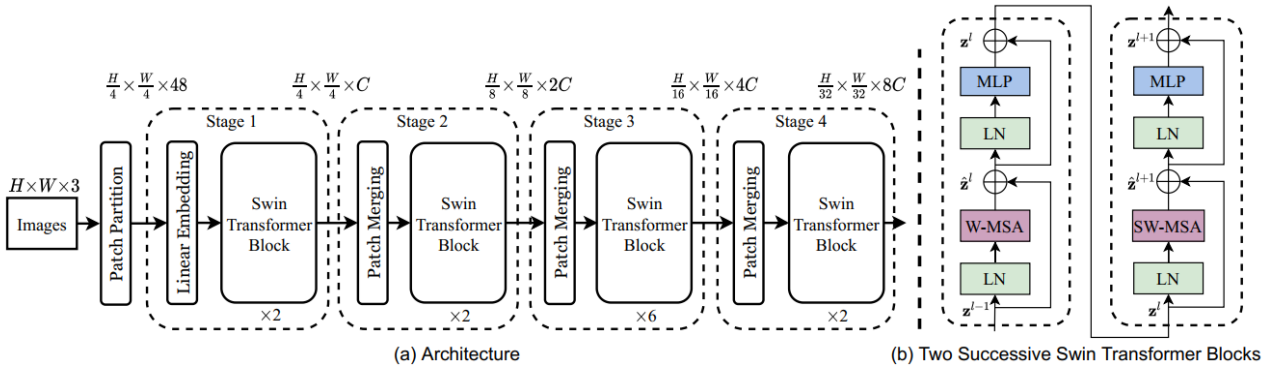

Figure 2. *The architecture of a Swin Transformer (Swin-T).*

A Swin Transformer block consists of W-MSA and SW-MSA modules, followed by a 2-layer MLP with GELU. A LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module.

Swin Transformer has a hierarchical representation with several Transformer blocks, and each block extracts different scale features. Low-level features like edges are extracted in shallower blocks, while high-level features like semantic dependencies among objects in deeper blocks. However, it only considers the attention for features of one layer, but ignores the combination of attention in different layers. In fact, human can understand images by paying attention the given images hierarchically to combine different scales instead of considering each scale separately. Therefore, global attention by incorporating the feature maps of different layers can extract more useful information based on their relationship.

Based on these studies, we propose the enhanced Swin Transformer, which combine the global feature output and the output obtained by global attention with the fusion of different scale feature maps from each stage of Swin Transformer.

3. Proposed Approach

3.1. Product Recognition System

We propose a product recognition system that consists of a two-stage pipeline with product detection and recognition (Figure 3). First, YOLOv11 [3] is utilized for the product-agnostic detection, which predicts bounding boxes for products. The image patch corresponding to each bounding box is then cropped and resized into 112 ×112. Second, the enhanced Swin Transformer is employed to extract feature embedding from this patch. Following that, we use K-NN (K-Nearest Neighbors) classifier for a similarity search between a feature embedding and feature embeddings computed on the reference database. This proposed approach combines both stages of product detection and recognition into a unified pipeline, leveraging the advantages of each method while avoiding the limitations imposed by single-stage detection models.

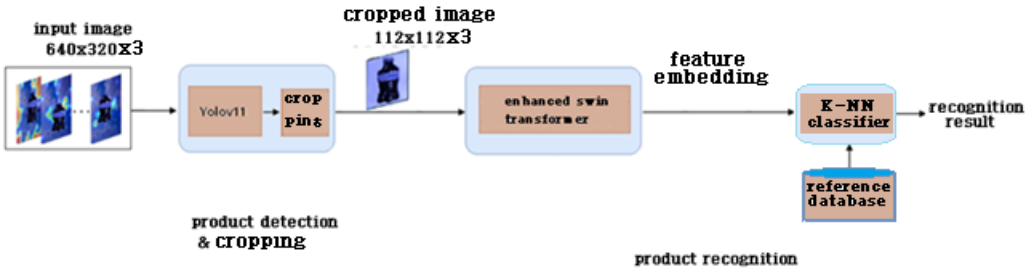

Figure 3. The proposed product recognition system.

3.2. Enhanced Swin Transformer

As mentioned above, the shallower blocks of original Swin Transformer extracts low-level features like colors or edges, while the deeper blocks extracts high-level features like semantic relationship between different objects. However, the attention is applied for the features of one layer and is not considered for different scale feature maps. Therefore, global attention by incorporating the feature maps of different layers can extract more useful information and improve generalization capability of the model.

Based on this study, we propose the enhanced Swin Transformer as shown in Figure 4. The proposed enhanced Swin Transformer consists of Swin Transformer backbone and GMSA. We use an original Swin Transformer backbone proposed in [7] to extract shared feature maps from different transformer blocks. Then the final output is obtained by combining the global feature output and the output obtained through the global attention module.

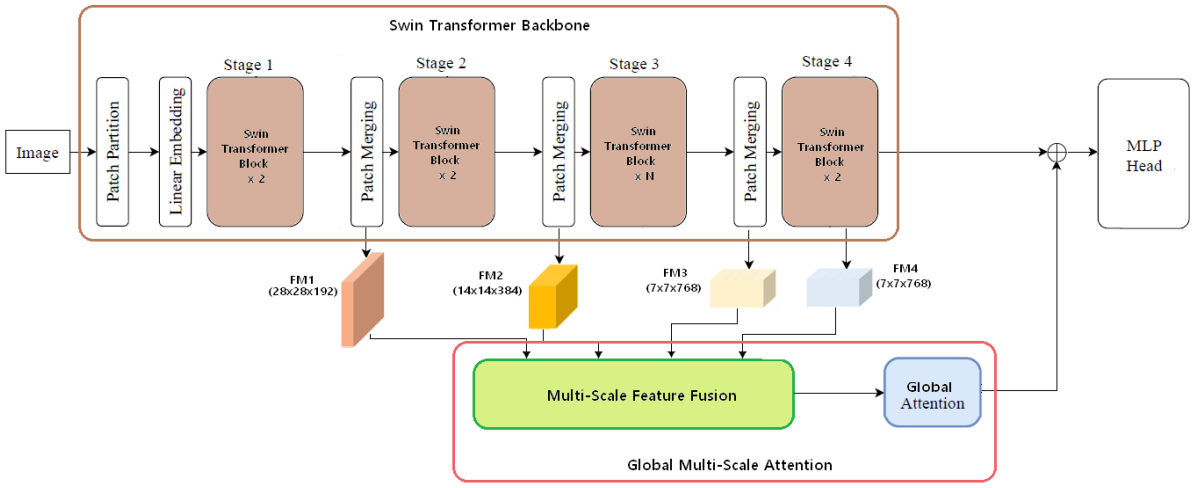

Figure 4. Diagram of the enhanced Swin Transformer.

With the shifted window partitioning approach, consecutive Swin Transformer blocks are computed as

$$\hat{z}^{l} = \text{W-MSA}\left(\text{LN}(z^{l-1})\right) + z^{l-1},$$

$$z^{l} = \text{MLP}\left(\text{LN}(\hat{z}^{l})\right) + \hat{z}^{l},$$

$$\hat{z}^{l+1} = \text{SW-MSA}\left(\text{LN}(z^{l})\right) + z^{l},$$

$$z^{l+1} = \text{MLP}\left(\text{LN}(\hat{z}^{l+1})\right) + \hat{z}^{l+1}, \quad (1)$$

where z^{l} and $\hat{z}^{l}$ denote the output features of the (S)W-MSA module and the MLP module for block l. And we set the final output of the backbone as Out_{swin}.

$$Out_{swin} = z^{4} \quad (2)$$

GMSA is composed of Multi-Scale Feature Fusion (MSFF) Module [9] and Global Attention Module (Figures 5, 6). The output of MSFF is obtained as follows:

$$f_1 = Conv\left(AvgPool(z^{1})\right),$$

$$f_2 = Conv\left(AvgPool(z^{2})\right),$$

$$f_3 = Conv(z^3),$$

$$f_4 = Conv(z^4),$$

$$f_{MSFF} = Concat(f_1, f_2, f_3, f_4), \quad (3)$$

where $Conv(\ ), AvgPool(\ ), Concat(\ )$ are the convolution layer with kernel size of 3×3 and average pooling layer, concatenation operation respectively.

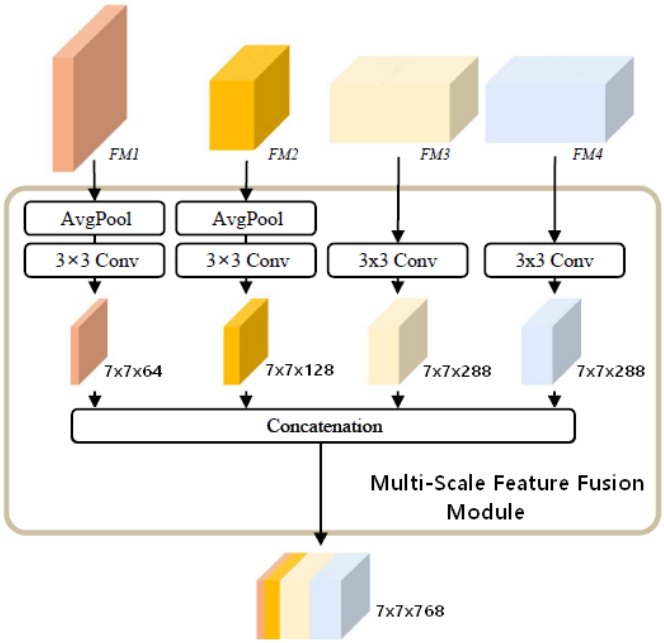

Figure 5. *Multi-Scale Feature Fusion Module.*

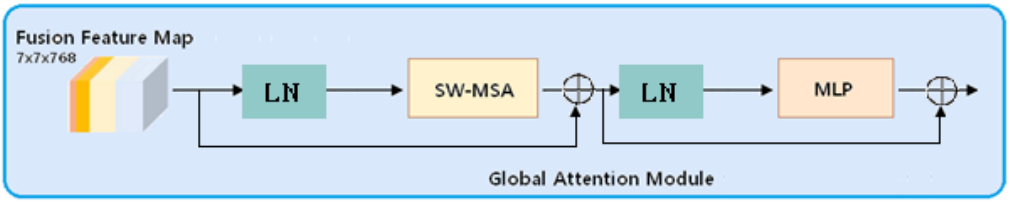

Figure 6. *Global Attention Module.*

Global Attention Module includes SW-MSA module, MLP and LN layer. A LN layer is applied before SW-MSA module and MLP, and a residual connection is applied after each module. The output of GMSA is obtained as follows:

$$f_{AM} = \text{SW-MSA}(\text{LN}(f_{MSFF})) + f_{MSFF}$$

$$Out_{GMSA} = \text{MLP}(\text{LN}(f_{AM})) + f_{AM} \quad (4)$$

The proposed transformer improves the generalization capabilities by combining attentions for each scale and global attention for the fused feature. With combination of the global feature output Out_{swin} and the GMSA output Out_{GMSA}, the final output is calculated by

$$Out = (1 - \alpha) \times Out_{swin} + \alpha \times Out_{GMSA} \quad (5)$$

where α is a weight coefficient and we set it as 0.4 through the experiments.

4. Experiments

4.1. Dataset

We construct a large-scale retail product dataset which contains 489 beverage bottles and 1840 food packages for training, and 30 products for validation.

There are around 30 photos for each product with diverse

pose variations as shown in Figure 7. In order to preprocess the data for training both stages of our product recognition system, we employ Labelme, a widely used segmentation tool for creating labeled images. This approach allows us to generate ground truth annotations for product detection and recognition tasks.

Figure 7. *Sample photos taken by a phone camera.*

For product-agnostic detection, we generate around 400K images for training and 40K images for validation by compositing background images and products with diverse pose variations, different scale and illumination changes as shown in Figure 8 to improve the model's ability to detect products. For product recognition, we generate a training dataset with 232K image patches which consists of 100 images patches for each product category by cropping product images with diverse changes. (Figure 9) We also generate a validation dataset with randomly chosen 20K image pairs from 30 products for validation.

Figure 8. *A sample image for product detection.*

Figure 9. Sample images for product recognition.

4.2. Experimental Setup

The experimental environment is a personal computer with NVIDIA GeForce RTX 3060. We use PyTorch framework for training and testing the models.

We choose YOLOv11s for the detection and use the same hyper-parameters as YOLOv11. For recognition, we use the AdamW optimizer and train the model for 30 epochs based on a cosine decay learning rate scheduler. The linear warm-up learning rate mechanism is adopted for the first 4 epochs. We set the batch size as 128 due to the limitation of GPU memory, and the base learning rate as 5×10^{-4}, a warm-up learning rate as 5×10^{-7}, a minimum learning rate as 5×10^{-6}, and a weight decay as 0.05. We employ horizontal and vertical flip, random erasing for data augmnetations.

4.3. Performance Evaluation

Table 1 shows the performance of the product-agnostic detection model on the validation dataset. We use the average precision (AP) metric for the intersection over union (IoU) thresholds set at 0.5, 0.5:0.95 is used for evaluation metric. The experimental result helps to obtain a comprehensive understanding of how well our detection model generalizes across diverse pose variations, illumination conditions, and object scales.

Table 1. Performance evaluation for the product detection model.

Metric	Result (%)
AP@[IoU=0.5]	99.486
AP@[IoU=0.5:0.95]	96.165

Table 2. Performance evaluation for the enhanced Swin Transformer according to different α values.

α	TAR@ [FAR=1e-4]	TAR@ [FAR=1e-3]	TAR@ [FAR=1e-2]	TAR@ [FAR=1e-1]	Accuracy (1-EER)
0.3	86.8	94.2	97.8	99.2	98.8
0.4	88.9	95.1	98.1	99.3	99.1

α	TAR@ [FAR=1e-4]	TAR@ [FAR=1e-3]	TAR@ [FAR=1e-2]	TAR@ [FAR=1e-1]	Accuracy (1-EER)
0.5	86.7	94.4	97.9	99.3	99.0
0.6	85.4	93.5	97.7	99.1	98.5

Table 3. *Performance evaluation for the product recognition methods.*

Model	TAR@ [FAR=1e-4]	TAR@ [FAR=1e-3]	TAR@ [FAR=1e-2]	TAR@ [FAR=1e-1]	Accuracy (1-EER)
Swin Transformer	86.7	93.9	97.7	99.2	98.8
The proposed model	88.9	95.1	98.1	99.3	99.1

For product recognition, we compare our enhanced Swin Transformer with global multi-scale attention to the original Swin Transformer on a validation dataset consisting of 20K product image pairs. To evaluate this comparison, we use accuracy (1-EER) and True Acceptance Rate (TAR) as evaluation metrics, which measure the model's performance across various false acceptance rates (FAR). Table 2 shows the performance evaluation for the enhanced Swin Transformer according to different α values of the equation (5) and 0.4 gives the best results. The results in Table 3 clearly demonstrate that the enhanced Swin Transformer improves generalization capabilities compared to the original Swin Transformer on our product recognition dataset. In practice, we use a threshold when FAR is equal to 1e-3 for true acceptance. The proposed method improves TAR by 1.2% when FAR is equal to 1e-3.

The proposed product recognition system using an enhanced Swin Transformer with global multi-scale attention demonstrates remarkable generalization capabilities and can handle boundless products without retraining the models.

5. Conclusions

In this paper, we proposed the effective product recognition system with two-stage pipeline consisting of Yolov11 based product-agnostic detection and product recognition based on the enhanced Swin Transformer to improve the product recognition performance on UVMs. We also proposed the enhanced Swin Transformer which combine the global feature output and the output obtained by global multi-scale attention with the fusion of different scale feature maps from each transformer block. Based on the experiments, the proposed product recognition system proves to be highly effective and suitable for various real-world scenarios. We implemented the proposed system on PC with RTX3060 and achieved a real-time speed for two cameras, but RTX3060 costs too expensive and is not suitable for the scenario which requires only one UVM. In the future, we plan to improve the product recognition performance by studying further vision transformers and implement a stand-alone system on Raspberry PI with Hailo NPU devices.

Abbreviations

UVM	Unmanned Vending Machine
ViT	Vision Transformer
YOLO	You Only Look Once
GMSA	Global Multi-Scale Attention
CNN	Convolutional Neural Network
IoU	Intersection of Union
TAR	True Acceptance Rate
FAR	False Acceptance Rate

Author Contributions

Kwang-Min Kim: Conceptualization, Investigation, Project administration, Software, Writing – original draft

Kwang-Uk Han: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft

Song-Jun Yun: Investigation, Formal Analysis, Resources, Visualization, Writing – review & editing

Hakho Hong: Data curation, Formal Analysis, Investigation, Validation, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bazarevsky V et al., “BlazePose: on-device real-time body pose tracking”, arXiv:2006.10204, 2020.

[2] Hong S et al., "Building unmanned store identification systems using YOLOv4 and Siamese network", Appl. Sci., Vol. 12, No. 8, p. 3826, 2022 https://doi.org/10.3390/app12083826

[3] Jocher G, Yolov11. https://github.com/ultralytics 2024.

[4] Li J et al., "BP-YOLO: a real-time product detection and shopping behaviors recognition model for intelligent unmanned vending machine", IEEE Access, Vol. 12, pp. 21038-21051, 2024 https://doi.org/10.1109/ACCESS.2024.3361675

[5] Liu L et al., "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, 2022 https://doi.org/10.1109/MCE.2021.3060722

[6] Liu Z et al., "Swin transformer v2: scaling up capacity and resolution", in IEEE/CVF Conf. on CVPR, pp. 11999-12009, 2022.

[7] Liu Z et al., "Swin transformer: hierarchical vision transformer using shifted windows", in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 10012-10022, 2021.

[8] Palanisamy B et al., "Transformers for vision: a survey on innovative methods for computer vision", IEEE Access, Vol. 13, pp. 95496-95523, 2025 https://doi.org/10.1109/ACCESS.2025.3571735

[9] Qin L et al., "SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation", arXiv:2308.11509v1, 2023.

[10] Santa B et al., "A comprehensive survey on computer vision based approaches for automatic identification of products in retail store", Image and Vision Computing, 86: 45-63, 2019.

[11] Sinha A et al., "An improved deep learning approach for product recognition on racks in retail stores", arXiv:2202.13081v1, 2022.

[12] Xia K et al., "An intelligent self-service vending system for smart retail", Sensors, Vol. 21, No. 10, p. 3560, 2021 https://doi.org/10.3390/s21103560

[13] Xu J et al., "Research on product detection and recognition methods for intelligent vending machines", Front. Neurosci., 17: 1288908, 2023 https://doi.org/10.3389/fnins.2023.1288908

[14] Wang C et al., "YOLOv7: Trainable bag-of-freebies sets new state-of-art for real-time object detectors", arXiv:2207.02696, 2022.

[15] Wei Y et al., "Deep learning for retail product recognition: challenges and techniques", Computational Intelligence and Neuroscience, Article ID: 8875910, 2020.