Two-Stage Product Detection and Classification for Retail Shelf Recognition

Chien-Wei Weng, Yuhong Li, Yingzhou Fang, and Ke Chen

CentraleSupélec, Université Paris-Saclay
MSc Data Sciences and Business Analytics
{chien-wei.weng, yuhong.li, yingzhou.fang, ke.chen}@student-cs.fr

Abstract. We present a two-stage deep learning pipeline for automated product recognition on retail shelves, combining YOLOv5 detection with ResNet-18 classification. The modular architecture enables independent optimization of detection and classification stages using different datasets: SKU-110K (11,762 shelf images) for detection and Grocery Store Dataset (5,125 images, 81 categories) for classification. Our detection stage achieves 88.9% mAP@0.5 with 89.6% precision on dense shelf environments (average 147 products per image). The classification stage reaches 78.3% test accuracy with 77.8% macro F1-score, identifying a circular misclassification pattern among seven visually similar vegetable classes. End-to-end integration processing 7,990 product regions demonstrates pipeline functionality while quantifying the impact of dataset domain mismatch (3.69% average confidence). The results validate the two-stage approach for retail product recognition and highlight the critical importance of dataset alignment for practical deployment.

Keywords: Product recognition · Object detection · Image classification · Retail automation · Deep learning

1 Introduction

Automated product recognition on retail shelves presents significant challenges for computer vision systems due to dense object arrangements, severe occlusion, and the need for fine-grained classification among visually similar products. Accurate automated recognition enables critical retail applications including inventory management, planogram compliance verification, and out-of-stock detection—addressing industry pain points that currently require manual inspection. With retail environments containing an average of 147 products per shelf image [1], traditional single-stage detection approaches struggle to balance localization accuracy with classification precision across diverse product categories.

Recent advances in deep learning have enabled two distinct approaches to this problem: single-stage detectors that perform detection and classification simultaneously [6,5], and two-stage pipelines that separate these tasks [7]. While single-stage methods offer computational efficiency, two-stage architectures provide modularity and the ability to optimize each component independently.

In this work, we develop a two-stage product recognition pipeline combining YOLOv5 [3] for detection with ResNet-18 [2] for classification. Our approach leverages two complementary datasets: SKU-110K [1] providing realistic shelf images with dense object annotations for detection training, and the Grocery Store Dataset [4] offering fine-grained category labels for classification. This modular design enables independent model optimization and dataset utilization without requiring unified labeling across both tasks.

Contributions. We present a functional two-stage pipeline achieving 88.9% detection mAP and 78.3% classification accuracy, identify a circular misclassification pattern among vegetable classes, and quantify dataset domain mismatch impact through integration analysis (3.69% average confidence).

The remainder of this paper is organized as follows: Section 2 reviews related work in retail product recognition. Section 3 describes our two-stage methodology. Section 4 presents experimental results for each stage and integration analysis. Section 5 concludes with findings and future directions.

2 Related Work

2.1 Object Detection Methods

Modern object detection has evolved through two main paradigms: two-stage and single-stage detectors. Two-stage methods like Faster R-CNN [7] first generate region proposals, then classify each region, achieving high accuracy at the cost of computational speed. Single-stage detectors such as YOLO [6] and SSD [5] perform detection and classification in one pass, prioritizing speed over precision. YOLOv5 [3] represents a modern evolution offering improved accuracy while maintaining real-time performance, making it suitable for dense retail environments.

2.2 Retail Product Recognition

Retail shelf recognition presents unique challenges including extreme object density, severe occlusion, and high intra-class visual similarity. Goldman et al. [1] introduced the SKU-110K dataset specifically addressing dense object detection with highly dense shelf images. Their work demonstrated that standard detection methods struggle with such density, motivating specialized approaches for retail environments.

For fine-grained product classification, Klasson et al. [4] proposed the Grocery Store Dataset with 81 hierarchically organized product categories captured under realistic retail conditions. Their work emphasized the importance of training on in-domain data for accurate product recognition, highlighting challenges in distinguishing visually similar items within the same product family.

2.3 Two-Stage Architectures

Two-stage pipelines separate detection and classification, enabling independent optimization of each component. This modularity is particularly valuable when training datasets have different characteristics—as in our case, where detection training uses dense shelf images while classification training uses isolated product images. The separation also allows leveraging different architectures optimized for each task: spatial localization for detection and fine-grained feature discrimination for classification.

Deep residual networks [2] have proven effective for image classification through their ability to train very deep models via skip connections. ResNet architectures pretrained on ImageNet provide strong feature extractors that transfer well to domain-specific tasks like grocery product recognition when fine-tuned appropriately.

Our work combines these approaches into a modular two-stage pipeline, using YOLOv5 for dense detection on shelf images and ResNet-18 for fine-grained classification, while explicitly analyzing the impact of dataset domain mismatch on end-to-end integration performance.

3 Methodology

3.1 Stage 1: Product Detection

Dataset. We use the SKU-110K dataset [1] containing 11,762 retail shelf images with 1.7 million bounding box annotations. We follow the standard train/validation/test split: 8,219 training images, 588 validation images, and 2,936 test images.

Model Architecture. We employ YOLOv5s, a single-stage detector optimized for real-time object detection. The model performs single-class detection with class label "object" rather than product-specific classification, focusing purely on spatial localization. This simplification is appropriate since fine-grained classification is handled by Stage 2.

Training Configuration. We fine-tune YOLOv5s pretrained on COCO using 3,000 training images (randomly sampled with fixed seed for reproducibility from the full 8,219-image training set) across three incremental batches of 1,000 images each. Training runs for 117 total epochs (69, 22, and 26 epochs per batch) with early stopping patience of 20 epochs. Training uses input resolution 640×640 and batch size 8. The model converges quickly due to strong COCO pretraining, with classification loss decreasing from 2.211 to 0.593 (73% reduction). The best model is selected based on validation mAP.

3.2 Stage 2: Product Classification

Dataset. We use the Grocery Store Dataset [4] containing 5,125 images across 81 fine-grained product categories organized hierarchically (e.g., Packages/Milk/Arla-Ecological-Medium-Fat-Milk). Images are captured under realistic retail lighting

and include natural backgrounds. We split the data into 2,640 training images and 2,485 test images.

Model Architecture. We employ ResNet-18 [2] pretrained on ImageNet, replacing the final fully connected layer to output 81 classes. All network parameters are trainable (11.2M parameters total), enabling full fine-tuning rather than just the classifier head.

Training Configuration. Following the training configuration from Klasson et al. [4], we initialize the final layer using Gaussian distribution $\mathcal{N}(0, 0.01)$ rather than PyTorch's default Kaiming initialization. We train for 30 epochs using SGD with momentum 0.9, weight decay 10^{-6}, and initial learning rate 0.001—optimizer settings adapted from their DenseNet-169 work. Input images are resized to 224×224 and normalized using ImageNet statistics. We apply random horizontal flips for data augmentation. Training uses cross-entropy loss and saves the model achieving best validation accuracy.

3.3 Stage 3: End-to-End Integration

We evaluate pipeline integration by applying the Stage 2 classifier to product regions extracted by the Stage 1 detector. We sample 50 images from the test set, from which the detector generates 7,990 cropped product regions (average 159.8 per image). Each crop is processed through the classification model to predict product categories and confidence scores. This integration test quantifies the impact of dataset domain mismatch: Stage 1 trains on diverse retail products (beverages, medicine, cosmetics) while Stage 2 trains on 81 specific grocery categories, creating a distribution gap we explicitly measure through confidence score analysis.

4 Experimental Results

4.1 Detection Performance (Stage 1)

Table 1 presents detection metrics on the SKU-110K validation set. Our YOLOv5s model achieves 88.9% mAP@0.5 with 89.6% precision and 81.8% recall. The high precision indicates accurate bounding box placement, while the lower recall reflects missed detections—approximately 18% of products remain undetected. The model shows strong localization with 56.2% mAP@0.5:0.95, confirming accurate spatial predictions across multiple IoU thresholds. Processing speed of 3.6ms per image enables real-time deployment.

The detector exhibits a 7% over-detection tendency, generating an average of 159.8 predictions per image compared to 147 ground truth objects. This behavior creates duplicate crops for some products, which the classification stage must handle robustly. Figure 1 shows detection outputs on sample shelf images, illustrating successful localization in dense, cluttered environments.

Table 1: Detection performance on SKU-110K validation set (588 images).

Metric	Value
Precision	89.6%
Recall	81.8%
mAP@0.5	88.9%
mAP@0.5:0.95	56.2%
Inference Speed	3.6 ms/image

Table 2: Classification performance on Grocery Store test set (2,485 images, 81 classes).

Metric	Value
Test Accuracy	78.31%
Macro F1-Score	77.75%
Training Accuracy	99.85%
Perfect Classes (F1=1.0)	5
Failed Classes (F1=0.0)	7

4.2 Classification Performance (Stage 2)

Table 2 summarizes classification results on the Grocery Store test set. The model achieves 78.31% test accuracy with 77.75% macro F1-score, demonstrating strong performance across most product categories. Training converges in 30 epochs with final training accuracy 99.85%, indicating 21.54 percentage points of overfitting gap.

Per-Class Analysis. Performance varies significantly across categories: 5 classes achieve perfect F1=1.0 scores (Papaya, Bravo-Orange-Juice, Arla-Mild-Vanilla-Yoghurt, Arla-Natural-Mild-Low-Fat-Yoghurt, Valio-Vanilla-Yoghurt), while 7 vegetable classes show complete failure (F1=0.0): Brown-Cap-Mushroom, Cabbage, Carrots, Cucumber, Garlic, Ginger, and Leek. Figure 2 shows training progression over 30 epochs.

Circular Misclassification Pattern. Analysis of the 7 failed classes reveals a striking circular misclassification pattern (Figure 3). These classes predominantly misclassify into each other: Brown-Cap-Mushroom → Cabbage (100%), Cabbage → Carrots (100%), Carrots → Cucumber (93%), and so forth, forming a closed loop. Crucially, 92% of misclassifications (173/188 samples) occur within this 7-class group, with only 8% confused with other categories. This suggests the model treats these vegetables as a continuous visual spectrum rather than discrete classes, likely due to shared visual features (elongated shapes, similar colors, vegetable context).

4.3 Integration Analysis (Stage 3)

Applying the classifier to 7,990 detection crops yields an average confidence of 3.69% (median 3.48%, std 0.98%) (Figure 4)—dramatically lower than typical classification confidence scores. Only 3 predictions (0.04%) exceed 10% confidence, all for Zucchini class.

This extreme low confidence quantifies the dataset domain mismatch impact: SKU-110K detection crops contain diverse retail products (Pepsi cans, medicine boxes, cosmetics) while the classifier trained exclusively on 81 grocery categories. The model appropriately exhibits uncertainty on out-of-distribution inputs rather than overconfidently assigning incorrect labels. Top predictions show bias toward cylindrical package categories (Oatghurt 14.2%, Yoghurt 11.6%), reflecting training distribution and visual similarity to cropped bottle shapes.

The integration successfully processes all regions at 165 images/second, demonstrating technical viability. However, practical deployment requires either unified dataset labeling or confidence thresholding to reject low-confidence predictions.

5 Conclusion

We have presented a two-stage deep learning pipeline for automated retail product recognition, combining YOLOv5 detection with ResNet-18 classification.

Key Contributions. Our work makes the following contributions:

- A functional two-stage pipeline achieving 88.9% detection mAP@0.5 and 78.3% classification accuracy, demonstrating the viability of modular architecture for retail product recognition.
- Identification of a circular misclassification pattern among seven vegetable classes, where 92% of errors occur within the group—providing insights into model failure modes and suggesting hierarchical classification approaches.
- Quantification of dataset domain mismatch impact through end-to-end integration, showing 3.69% average confidence when applying the classifier to detection outputs—highlighting the critical importance of dataset alignment for practical deployment.
- Comprehensive evaluation methodology including per-class F1 analysis, confusion matrices for failed classes, and processing speed benchmarks (165 images/second).

Limitations and Future Work. The primary limitation is dataset domain mismatch. Future work should: (1) collect unified labels across datasets, (2) implement confidence thresholding for out-of-distribution detection, and (3) explore hierarchical classification for visually similar product families.

Our results validate two-stage architectures for retail product recognition while providing a baseline for dataset alignment strategies.

Code and trained models are available at `https://github.com/wengchienwei/retail-shelf-product-recognition`.

References

1. Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., Hassner, T.: Precise Detection in Densely Packed Scenes. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5227–5236. IEEE, Long Beach (2019). https://doi.org/10.1109/CVPR.2019.00537
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Las Vegas (2016). https://doi.org/10.1109/CVPR.2016.90
3. Jocher, G., et al.: YOLOv5. GitHub repository. https://github.com/ultralytics/yolov5 (2020). Accessed 24 Jan 2026
4. Klasson, M., Zhang, C., Kjellström, H.: A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 491–500. IEEE, Waikoloa (2019). https://doi.org/10.1109/WACV.2019.00058
5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. IEEE, Las Vegas (2016). https://doi.org/10.1109/CVPR.2016.91
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149. IEEE (2017). https://doi.org/10.1109/TPAMI.2016.2577031

Fig. 1: Detection results on SKU-110K test images showing successful localization in dense shelf environments. Red boxes indicate predicted bounding boxes. The model handles severe occlusion and high object density (up to 576 products per image).

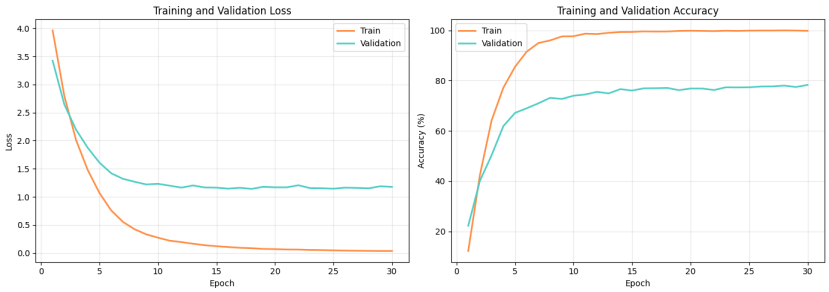

Fig. 2: Training and validation curves over 30 epochs for Stage 2 classification. The model achieves best validation accuracy of 78.31% at epoch 30, with a 21.54 percentage point gap between training (99.85%) and validation performance indicating overfitting.

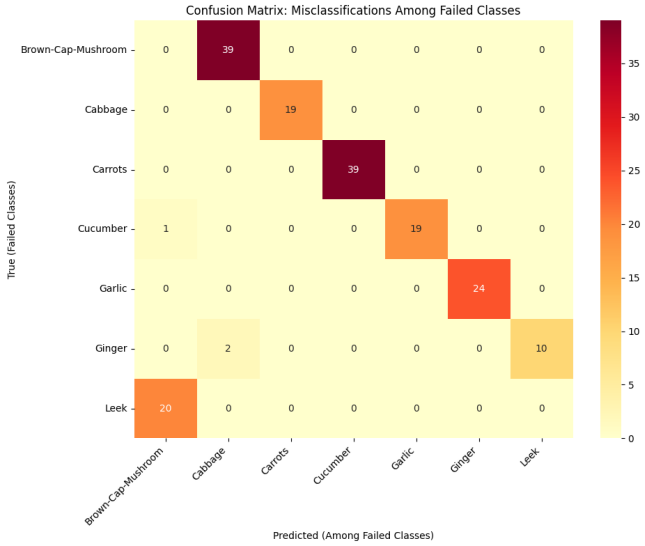

Fig. 3: Confusion matrix for 7 failed vegetable classes showing circular misclassification pattern. 92% of errors (173/188 samples) occur within this group: Brown-Cap-Mushroom → Cabbage → Carrots → Cucumber → Garlic → Ginger → Leek → Brown-Cap-Mushroom, forming a closed loop.

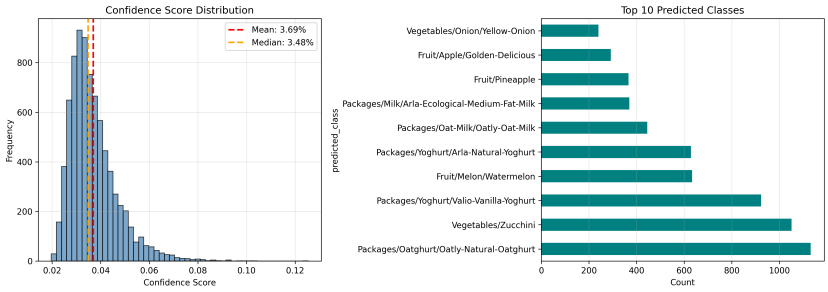

Fig. 4: Confidence score distribution for Stage 3 integration (7,990 predictions). Average confidence is 3.69% with 90.6% of predictions falling in 1–5% range, quantifying the severe dataset domain mismatch between SKU-110K detection crops and Grocery Store classification training.