Jurnal Pendidikan, Sains dan Teknologi
Vol 13, No 01, 2026 pages 146 - 157
e-ISSN : 2655-3392
https://doi.org/10.47668/edusaintek.v13i1.2154

DETECTION OF VISUALLY SIMILAR RETAIL PRODUCTS WITH SIZE VARIANTS: A COMPARISON OF YOLOv12 AND FASTER R-CNN

Ganendra Raditya Putra Satriawan[1]*, Dhomas Hatta Fudholi[2]
[1,2] Universitas Islam Indonesia, Indonesia
*e-mail korespondensi: ganendraditya@gmail.com

Abstrak: Otomasi ritel menghadirkan berbagai tantangan dalam aplikasi visi komputer, khususnya pada pengenalan produk *fine-grained*, di mana satu merek sering kali hadir dalam beberapa varian ukuran atau gramasi dengan desain kemasan yang hampir identik dan tersusun rapat di rak toko. Penelitian ini menganalisis kinerja dasar dari dua paradigma deteksi objek yang banyak diadopsi: YOLOv12 satu tahap dan Faster R-CNN dua tahap. Kedua arsitektur dievaluasi menggunakan dataset ritel khusus yang terdiri dari 408 gambar dengan 47 kelas, yang dikurasi untuk menekankan pengenalan varian ukuran pada kondisi rak nyata. Dengan konfigurasi pelatihan dan inferensi standar, hasil eksperimen menunjukkan bahwa dalam keluarga YOLOv12, varian *extra-large* (YOLOv12x) mencapai skor mAP50:95 sebesar 0,762 dengan *False Negative Rate* (FNR) 14,16%, sedangkan varian terkecil *nano* (YOLOv12n) mencatat mAP50:95 sebesar 0,544 dengan FNR 30,02%. Untuk paradigma dua tahap, Faster R-CNN dengan *backbone* ResNet50 dan ResNet50v2 masing-masing mencapai mAP50:95 sebesar 0,698 dan 0,709, dengan nilai FNR terkait 36,72% dan 34,78%. Hasil ini menunjukkan bahwa pengenalan produk ritel *fine-grained* masih merupakan tugas yang menantang, terutama pada kemasan yang sangat mirip dengan perbedaan ukuran yang relatif kecil. Temuan ini memberikan tolok ukur empiris untuk pengembangan sistem deteksi hibrida atau sistem yang ditingkatkan di masa mendatang.

Kata Kunci: *Produk Retail Fine-Grained, Deteksi Objek, YOLOv12, Faster R-CNN, Perbandingan Model*.

Abstract: Retail automation presents various challenges in computer vision applications, particularly in fine-grained product recognition, where a single brand often appears in multiple size or grammage variants with nearly identical packaging designs and is densely arranged on store shelves. This study analyzes the baseline performance of two widely adopted object detection paradigms: the single-stage YOLOv12 and the two-stage Faster R-CNN. Both architectures are evaluated on a custom retail dataset comprising 408 images with 47 classes, curated to emphasize size-variant recognition under real-world shelf conditions. Using standard training and inference configurations, the experimental results show that within the YOLOv12 family, the extra-large variant (YOLOv12x) achieves a mAP50:95 score of 0.762 with a False Negative Rate (FNR) of 14.16%, while the smallest nano variant (YOLOv12n) records a mAP50:95 of 0.544 with an FNR of 30.02%. For the two-stage paradigm, Faster R-CNN with ResNet50 and ResNet50v2 backbones achieves mAP50:95 scores of 0.698 and 0.709, with corresponding FNR values of 36.72% and 34.78%, respectively. These results indicate that fine-grained retail product recognition remains a challenging task, particularly for visually similar packaging designs with different size variants. The findings provide an empirical baseline for future hybrid or enhanced detection systems.

Keywords: *Fine-Grained Retail Product, Object Detection, YOLOv12, Faster R-CNN, Model Comparison*

Histori Naskah
Diserahkan: 23-07-2026
Direvisi: 26-01-2026
Diterima: 21-12-2025

INTRODUCTION

The global retail landscape is currently undergoing a structural metamorphosis toward digital automation to enhance operational efficiency and consumer shopping experiences

(Grewal et al., 2017). In response to significant challenges such as inefficient queue management and poor demand forecasting, retailers are increasingly integrating artificial intelligence and computer vision technologies (Hossam et al., 2024). These technologies are central to the realization of smart retail systems, which include automated checkout and real-time stock tracking, as they are considered more reliable and time-saving than manual operations (Wei et al., 2020). The fundamental objective of these systems is to achieve high-fidelity recognition of thousands of stock-keeping units (SKUs) to ensure seamless retail execution (Wang et al., 2022).

However, unlike general object detection, the retail environment presents a specific fine-grained challenge (Baz et al., 2016). Many product classes are visually almost identical in terms of shape, color, and texture, which often leads to significant classification errors (Baz et al., 2016). This issue is particularly evident in products from the same brand that share identical packaging designs but differ only in minute details such as sizes (Srivastava & Kumar, 2021). For instance, liquid products like bottled water may have identical visual appearances across different volume variants, making them extremely difficult to distinguish. Furthermore, retail product classification is highly complex because these similar items are often densely packed on shelves, resulting in extremely low inter-class variance (Srivastava, 2020).

To meet the strict latency requirements of smart retail applications, single-stage detectors such as the You Only Look Once (YOLO) architecture have become the industry standard due to their superior inference speed and efficiency (Fudholi et al., 2024). Recent studies in the retail sector have consistently utilized the YOLO framework to address complex scenarios involving small-scaled products and on-shelf availability (Fudholi et al., 2024). Furthermore, the evolution of this architecture has led to more robust variants, such as YOLOv11, which incorporates advanced feature extraction to handle multi-scale targets and dense occlusion in automated supermarkets (Hou & Huang, 2025). To maintain high accuracy in densely packed environments, hybrid approaches combining YOLO with other networks have also been proposed to leverage both global context and precise localization (Yazdanjouei et al., 2025). Building upon these architectural milestones, this study implements the latest iteration, YOLOv12, as the core detector. YOLOv12 introduces significant improvements over its predecessors by integrating attention-centric mechanisms, specifically Area Attention and Residual Efficient Layer Aggregation Networks (R-ELAN), which theoretically enhance the model's discriminative capability in fine-grained tasks (Tian et al., 2026).

Conversely, distinct from the single-stage paradigm, two-stage detectors represented by Faster R-CNN have long been established as the gold standard for tasks requiring high-precision localization (Ren et al., 2016). Unlike YOLO which predicts bounding boxes directly, Faster R-CNN employs a Region Proposal Network (RPN) to generate candidate regions before classification. Literature suggests that this two-stage learning framework is essential for fine-grained tasks to effectively localize discriminative regions where visual differences are subtle (He et al., 2017). This structural advantage keeps Faster R-CNN relevant in modern industrial applications despite the speed of newer models (Rane, 2023). Moreover, recent developments have led to significantly enhanced versions of this architecture. Specifically, this study utilizes an improved Faster R-CNN iteration with a ResNet-50-FPN backbone, which incorporates modern training recipes and architectural refinements, such as updated box heads and improved FPN structures proposed by Li et al. (2021) in their benchmarking of vision transformer transfer learning. Empirical evidence from the retail sector confirms its continued efficacy; it has been successfully implemented to enable smart checkout systems that eliminate manual scanning (Shailaja, 2020). Furthermore, a recent 2024 study on self-checkout cashiers explicitly utilizes Faster R-CNN to recognize grocery products in real-time, prioritizing its robust recognition capabilities over speed (Lauro et al., 2024).

In recent object detection tasks, a wide range of architectures have been proposed, including SSD, RetinaNet, EfficientDet, and transformer-based detectors. Nevertheless, YOLO and Faster R-CNN remain two of the most widely adopted and well-established detection frameworks in deployment. YOLO represents the dominant single-stage paradigm due to its real-time performance and deployment efficiency, while Faster R-CNN is widely regarded as a strong representative of the two-stage paradigm with proven localization accuracy. While each paradigm offers distinct advantages, their inherent trade-offs - speed versus localization precision - become critical in fine-grained retail scenarios where subtle visual differences must be distinguished under real-time constraints. Therefore, selecting both YOLOv12 (as the latest single-stage architecture) and Faster R-CNN (as a robust two-stage baseline) allows this study to systematically compare the two dominant detection paradigms and provide a balanced empirical baseline for recognizing visually similar retail product size variants.

Despite the extensive implementation of object detection in retail, research specifically targeting the distinction of fine-grained products especially with size variants remains limited. The challenge becomes significantly more complex when products share identical packaging designs across different volumes. Srivastava & Kumar (2021) emphasize this difficulty, noting that deciphering such size-based variants using standard computer vision algorithms is often deemed impractical due to extreme visual similarities. Addressing this gap, this study conducts an analysis of YOLOv12 and Faster R-CNN to evaluate their performance in recognizing fine-grained retail products with highly similar designs but different size attributes. The objective is to provide empirical evidence on the strengths and limitations of both detection paradigms in this challenging retail scenario and to establish a reference on similar recognition tasks.

METHOD

This study employs a structured experimental workflow designed to evaluate the performance of the YOLOv12 and Faster R-CNN architecture for fine-grained retail product detection. The methodology follows a sequential process consisting of four main stages. The overall workflow is visualized in Figure 1.

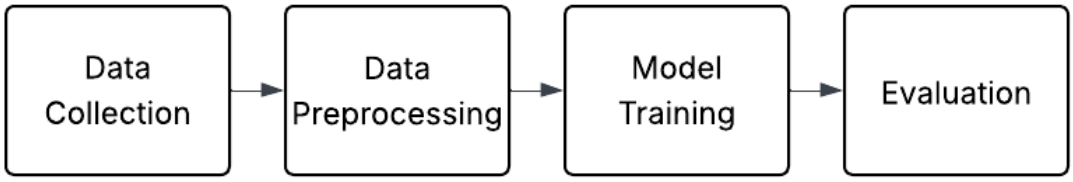

Figure 1. Research methodology overview

Data Collection

The dataset used in this research was constructed from observations in Indonesian retail stores, specifically targeting products with visually similar packaging designs and different size or grammage variants. The data collection focused on consumer products commonly found in Indonesian local supermarkets. Images were captured through field observations to reflect diverse shelf arrangements and lighting conditions encountered in practical retail environments.

Figure 2. Data example shows size variants of Dancow 1+ Madu with similar designs

The data was collected and curated by the researchers themselves. The dataset is not publicly available due to confidentiality restrictions. The dataset consists of 408 images containing a total of 2,614 annotated object instances across 47 fine-grained classes, which are further grouped into 13 major product families. Bounding boxes were drawn tightly around each product package, including the grammage label when visible, to ensure that size-related discriminative cues are captured. The detail of the dataset is shown in Table 1.

Table 1. List of product classes

Product Family	Class Name	Instances
DANCOW 1+ Madu	DANCOW 1+ Madu 1kg	52
	DANCOW 1+ Madu 750g	72
	DANCOW 1+ Madu 350g	99
	DANCOW 1+ Madu 120g	105
DANCOW 1+ Vanilla	DANCOW 1+ Van 1kg	49
	DANCOW 1+ Van 750g	105
	DANCOW 1+ Van 350g	107
	DANCOW 1+ Van 120g	42
DANCOW 3+ Madu	DANCOW 3+ Madu 1kg	85
	DANCOW 3+ Madu 750g	67
	DANCOW 3+ Madu 350g	85
DANCOW 3+ Vanilla	DANCOW 3+ Van 1kg	90
	DANCOW 3+ Van 750g	74
	DANCOW 3+ Van 350g	92
DANCOW 5+ Madu	DANCOW 5+ Madu 1kg	39
	DANCOW 5+ Madu 750g	94

	DANCOW 5+ Madu 350g	144
LACTOGEN 1 Happynutri	LACTOGEN 1 Happynutri 1kg	34
	LACTOGEN 1 Happynutri 735g	31
	LACTOGEN 1 Happynutri 350g	38
	LACTOGEN 1 Happynutri 180g	34
LACTOGEN 2 Happynutri	LACTOGEN 2 Happynutri 1kg	26
	LACTOGEN 2 Happynutri 735g	34
	LACTOGEN 2 Happynutri 350g	55
	LACTOGEN 2 Happynutri 180g	35
LACTOGEN PRO 0-6 mo	LACTOGEN PRO 0-6 mo 1kg	54
	LACTOGEN PRO 0-6 mo 735g	69
	LACTOGEN PRO 0-6 mo 350g	47
	LACTOGEN PRO 0-6 mo 180g	56
LACTOGEN PRO 6-12 mo	LACTOGEN PRO 6-12 mo 1kg	57
	LACTOGEN PRO 6-12 mo 735g	43
	LACTOGEN PRO 6-12 mo 350g	44
	LACTOGEN PRO 6-12 mo 180g	17
LACTOGROW 3 Happynutri Honey	LACTOGROW 3 Honey 1kg	29
	LACTOGROW 3 Honey 735g	36
	LACTOGROW 3 Honey 350g	36
	LACTOGROW 3 Honey 145g	26
LACTOGROW 3 Happynutri Vanilla	LACTOGROW 3 Van 1kg	14
	LACTOGROW 3 Van 735g	17
	LACTOGROW 3 Van 350g	29

LACTOGROW PRO 1+ Honey	LACTOGROW PRO 1+ Honey 1kg	14
	LACTOGROW PRO 1+ Honey 735g	25
	LACTOGROW PRO 1+ Honey 350g	84
	LACTOGROW PRO 1+ Honey 145g	36
LACTOGROW PRO 1+ Vanilla	LACTOGROW PRO 1+ Van 1kg	23
	LACTOGROW PRO 1+ Van 735g	25
	LACTOGROW PRO 1+ Van 350g	32

Data Preprocessing

To ensure data integrity and compatibility with both detection architectures, a structured preprocessing pipeline was applied. All images were manually annotated using bounding box annotations through the Roboflow platform. Each product instance was labeled according to its category, which includes brand, product variant, and size or grammage information. Since YOLO and Faster R-CNN require different annotation formats, the annotated dataset was exported into two separate label representations. For YOLO training, the annotations were converted into text-based (txt) format files containing normalized bounding box coordinates. For Faster R-CNN training, the annotations were exported in JSON format following standard object detection annotation structures. This dual-format preparation ensures that both detection frameworks can be trained using identical image content while preserving their respective annotation requirements.

To prevent data leakage, a group-based splitting strategy was employed (Rumala, 2023). Images were grouped based on their specific shelf scenes, ensuring that all images from a single shelf arrangement were confined entirely to either the training or testing partition. This approach follows the definition of data leakage where visually similar contexts appearing across training and testing sets can lead to overly optimistic evaluation results (Ramos et al., 2025). As a result, the model is evaluated on its ability to generalize to unseen scenes rather than memorizing background features.

Based on this strategy, the dataset was partitioned into a training set comprising 273 images (66.91%) with 1,579 bounding boxes, and a testing set comprising 135 images (33.09%) with 1,035 bounding boxes.

Model Training

The model training stage adopted a benchmarking of two distinct architectures: YOLOv12 and Faster R-CNN. This study evaluates the performance of various model configurations. To ensure robust generalization and prevent overfitting, early stopping was applied across all experiments (Hussein & Shareef, 2024). All YOLOv12 nano (n), small (s), medium (m), large (l), and extra-large (x) variants were trained using the default training presets provided by the Ultralytics framework, following the practice in the Ultralytics documentation (Jocher et al., 2023). The training configuration is summarized in Table 2.

Table 2. YOLOv12 configuration

Hyperparameter	Value
Training Epochs	100
Early stop patience	20
Batch size	16
Image size	640
Optimizer	auto
LR scheduler	auto
Warmup	auto

Conversely, Faster R-CNN was evaluated using two backbone architectures, ResNet50 and ResNet50v2. The overall training setup was mostly set automatically and also follows some parts from the reference training strategy described by Wu et al. (2019). Stochastic gradient descent (SGD) is employed as the primary optimizer together with a step-based learning rate scheduler. While several parameters were adapted to accommodate GPU memory constraints and dataset scale, the training configuration remains consistent with the recommended Faster R-CNN training protocol. The training duration was extended to 150 epochs to ensure sufficient convergence under limited data conditions. The detailed training configuration is provided in Table 3.

Table 3. Faster R-CNN configuration

Hyperparameter	Value
Training Epochs	150
Early stop patience	25
Batch size	2
Image size	auto
Optimizer	SGD
LR scheduler	StepLR
Learning rate	0.0005
Momentum	0.9
Weight decay	0.0005

Evaluation

The performance of the trained models was evaluated to benchmark the discriminative capabilities of the single-stage against the two-stage architecture performance. Detection performance is assessed using Mean Average Precision (mAP), specifically mAP50:95, following the standard COCO Primary Challenge protocol, and is complemented by global precision and recall to reflect practical detection behavior. In addition to detection performance, fine-grained classification errors are analyzed using class-wise confusion matrices to examine inter-variant misclassification patterns among visually similar product grammage categories. The confusion matrices are constructed by matching each ground-truth

object to the predicted bounding box with the highest Intersection-over-Union (IoU), where predictions are matched to ground-truth annotations based on IoU with a threshold of 0.5. Based on the resulting confusion matrices, classification failures are quantified using the False Negative Rate (FNR), also referred to as the Miss Rate (MR). Following the formulation from Jiang et al. (2025), the miss rate is defined as the ratio between the number of false negatives and the total number of actual positive instances. In this study, false negatives correspond to ground-truth objects that are either misclassified or not detected, and FNR is computed directly from the confusion matrix as the proportion of missed instances over the total number of ground-truth annotations.

RESULTS AND DISCUSSIONS

Experimental Results

In this section, we present a comparative performance analysis of the YOLOv12 architecture (n, s, m, l, and x variants) and the Faster R-CNN architecture with ResNet50 v1 and ResNet50v2 backbones. All models were evaluated on a dedicated test set consisting of 135 retail shelf images containing visually similar product variants. The comprehensive quantitative results, including mAP50:95, precision, and recall, are summarized in Table 4.

Table 4. Model performance comparison on fine-grained retail dataset

Architecture	Model Variant	mAP50:95	Precision	Recall
YOLOv12	Nano (n)	0.544	0.5433	0.6966
	Small (s)	0.692	0.6544	0.7903
	Medium (m)	0.760	0.6928	0.8454
	Large (l)	0.747	0.7084	0.8522
	Extra-large (x)	0.762	0.7118	0.8638
Faster R-CNN	ResNet50 (v1)	0.698	0.3377	0.9121
	ResNet50v2 (v2)	0.709	0.3568	0.8918

The confusion matrices of the YOLOv12 variants (n, s, m, l, and x) are presented in Figures 3-7, while the Faster R-CNN models with ResNet50 v1 and ResNet50 v2 backbones are shown in Figure 8 and Figure 9.

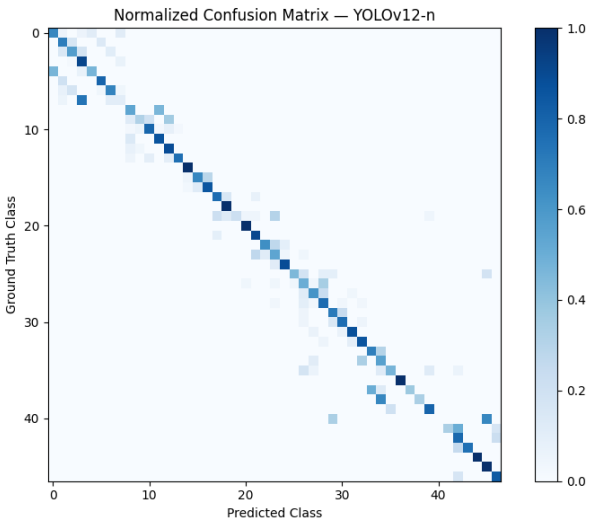

Figure 3. YOLOv12n confusion matrix

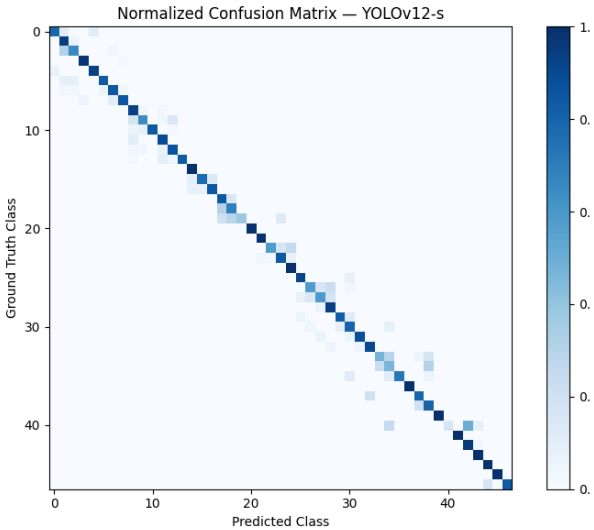

Figure 4. YOLOv12s confusion matrix

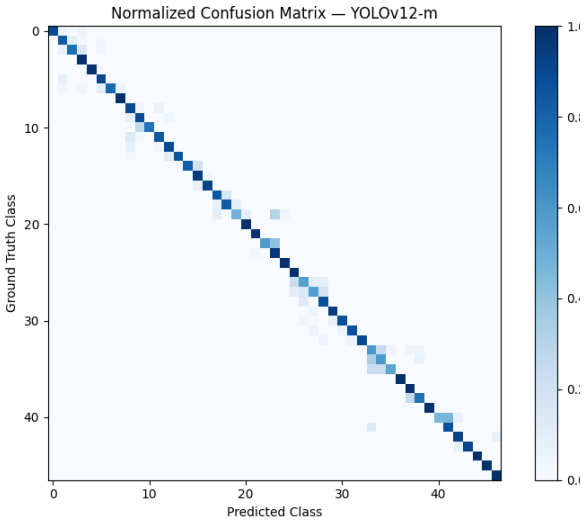

Figure 5. YOLOv12m confusion matrix

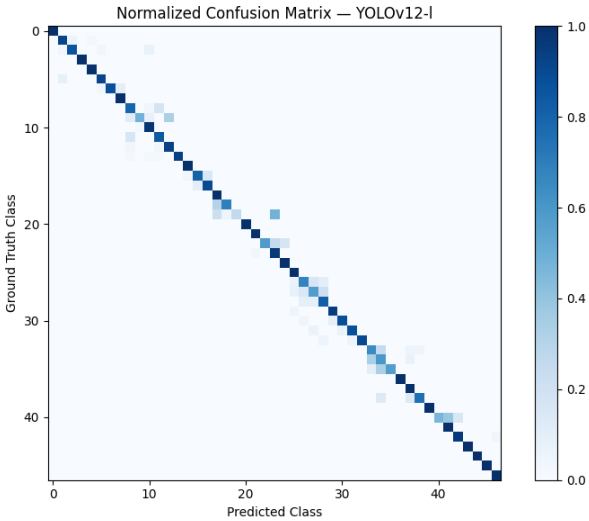

Figure 6. YOLOv12l confusion matrix

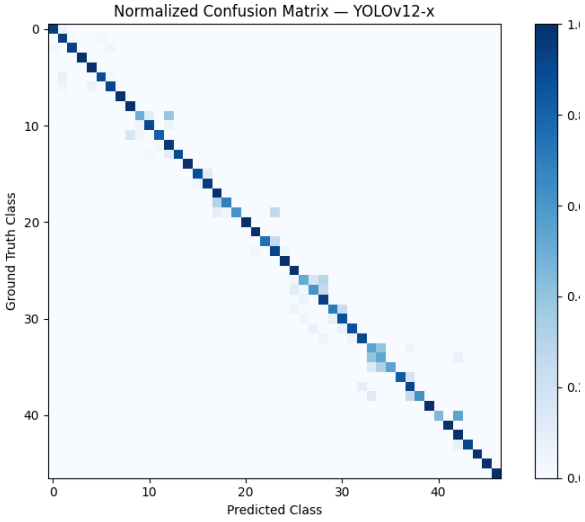

Figure 7. YOLOv12x confusion matrix

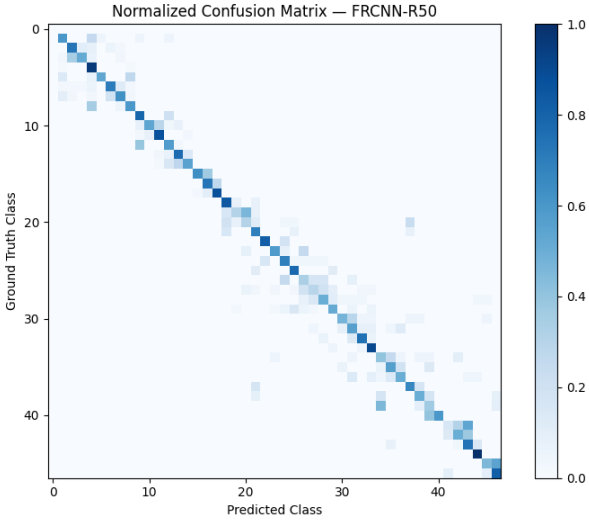

Figure 8. Faster R-CNN ResNet50 confusion matrix

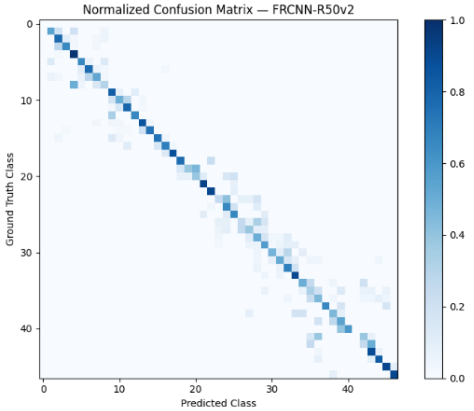

Figure 9. Faster R-CNN ResNet50v2 confusion matrix

The False Negative Rate (FNR) of YOLO and Faster R-CNN models is computed from the previous confusion matrices with IoU threshold of 0.5. The FNR results of each model are shown in the following Table 5.

Table 5. False Negative Rate of each model

Architecture	Model Variant	FNR
YOLOv12	Nano (n)	30.02%
	Small (s)	18.71%
	Medium (m)	15.87%
	Large (l)	14.80%
	Extra-large (x)	14.16%
Faster R-CNN	ResNet50 (v1)	36.72%
	ResNet50v2 (v2)	34.78%

Discussions

The experimental results demonstrate clear differences in detection performance across the evaluated model architecture. Among all configurations, the YOLOv12 family shows a consistent scaling trend, where increasing model capacity leads to higher detection accuracy. The YOLOv12n variant achieves a mAP50:95 of 0.544, while the YOLOv12x variant reaches 0.762, corresponding to an absolute improvement of 21.8 percentage points. In comparison, the Faster R-CNN models achieve mAP50:95 values of 0.698 for the ResNet50 v1 backbone and 0.709 for the ResNet50 v2 backbone, indicating that single-stage detectors with larger model capacity provide a balance between accuracy and efficiency for dense retail product recognition.

Beyond aggregate detection accuracy, the confusion matrix analysis and miss rate evaluation provide further insights into model robustness. The False Negative Rate (FNR) decreases consistently with increasing YOLOv12 model capacity, from 30.02% for YOLOv12n to 14.16% for YOLOv12x. In contrast, the Faster R-CNN models exhibit higher miss rates, with FNR values of 36.72% and 34.78% for the ResNet50 v1 and v2 backbones, respectively.

Nevertheless, the confusion matrix and miss rate evaluation indicate that fine-grained retail product recognition remains a challenging task, particularly for visually similar size variants. The observed misclassification patterns are likely influenced by a combination of factors, including class imbalance, limited training samples for specific grammage categories, and the inherent visual similarity of packaging designs that differ primarily in scale or minor textual cues.

A closer examination of the confusion matrices reveals that certain classes with very few training samples, such as LACTOGROW PRO 1+ Honey 1kg (only 14 instances), experienced near-total detection failure in smaller models, illustrating the impact of data imbalance. While larger YOLOv12 variants mitigated such extreme failures, the persistent confusion between adjacent grammage sizes (e.g., 735g vs 350g) suggests that increasing model capacity alone is insufficient. Regarding the two-stage detectors, the incremental mAP improvement from ResNet50 v1 to v2 (1.1%) is accompanied by a reduction in FNR (from 36.72% to 34.78%) and a slight gain in precision (0.3377 to 0.3568), indicating consistent but modest benefits. Finally, due to computational constraints, all models were trained once with fixed random seeds; therefore, the observed performance trends should be interpreted as observational rather than statistically validated. These limitations should be addressed in future work through multiple training runs and confidence interval reporting.

CONCLUSION

This study presents a benchmark evaluation of single-stage YOLOv12 and two-stage Faster R-CNN architectures for fine-grained retail product recognition. Within the YOLOv12 family, the experimental results demonstrate a performance scaling trend, where increasing model capacity leads to improved detection accuracy. The YOLOv12x variant achieves the highest performance with a mAP50:95 of 0.762, corresponding to an absolute improvement of 21.8 percentage points over the YOLOv12n variant. This trend is further supported by the miss rate analysis, where the False Negative Rate (FNR) decreases from 30.02% for YOLOv12n to 14.16% for YOLOv12x.

For the two-stage paradigm, the Faster R-CNN models with ResNet50 v1 and ResNet50v2 backbones exhibit consistent detection performance, with mAP50:95 values of 0.698 and 0.709, respectively. The corresponding miss rates of 36.72% (R50) and 34.78% (R50v2) indicate that the architectural refinements introduced in the v2 backbone provide incremental improvements under the training configuration in this study.

Moreover, the results indicate that fine-grained retail product recognition remains a challenging task. The observed model performance is likely influenced by dataset characteristics, including class imbalance and limited training samples for certain product variants, as well as the strong visual similarity between packaging designs. These observations suggest that increasing model capacity alone may not be sufficient to fully resolve fine-grained visual ambiguities under the current environment. The complementary strengths observed—YOLOv12x's efficiency and accuracy versus Faster R-CNN's stable recall—suggest that hybrid designs could benefit from combining both paradigms, e.g., using a two-stage detector as a verifier for ambiguous predictions.

Future research directions include expanding the dataset to provide more balanced and diverse training samples, exploring improved data augmentation and class rebalancing strategies. In addition, further optimization of preprocessing, modelling and post-processing configurations may improve detection robustness and reliability in real-world retail environments.

REFERENCES

Baz, I., Yoruk, E., & Cetin, M. (2016). Context-aware hybrid classification system for fine-grained retail product recognition. *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 1–5. https://doi.org/10.1109/IVMSPW.2016.7528213

Fudholi, D. H., Kurniawardhani, A., Andaru, G. I., Alhanafi, A. A., & Najmudin, N. (2024). YOLO-based Small-scaled Model for On-Shelf Availability in Retail. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *8*(2), 265–271. https://doi.org/10.29207/resti.v8i2.5600

Grewal, D., Roggeveen, A. L., & Nordfält, J. (2017). The Future of Retailing. *Journal of Retailing*, *93*(1), 1–6. https://doi.org/10.1016/j.jretai.2016.12.008

He, X., Peng, Y., & Zhao, J. (2017). Fine-grained Discriminative Localization via Saliency-guided Faster R-CNN. *Proceedings of the 25th ACM International Conference on Multimedia*, 627–635. https://doi.org/10.1145/3123266.3123319

Hossam, A., Ramadan, A., Magdy, M., Abdelwahab, R., Ashraf, S., & Mohamed, Z. (2024). Revolutionizing Retail Analytics: Advancing Inventory and Customer Insight with AI. *2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, 64–69. https://doi.org/10.1109/ICMISI61517.2024.10580424

Hou, P., & Huang, S. (2025). BCSM-YOLO: An Improved Product Package Recognition Algorithm for Automated Retail Stores Based on YOLOv11. *IEEE Access*, *13*, 139665–139679. https://doi.org/10.1109/ACCESS.2025.3595175

Hussein, B. M., & Shareef, S. M. (2024). An Empirical Study on the Correlation between Early Stopping Patience and Epochs in Deep Learning. *ITM Web of Conferences*, *64*, 01003. https://doi.org/10.1051/itmconf/20246401003

Jiang, B., Wang, J., Ren, G., Zhi, M., Yu, Z., Zhang, Y., Ren, P., & Jia, S. (2025). Research on pedestrian detection method based on multispectral intermediate fusion using YOLOv7. *Scientific Reports*, *15*(1), 16851. https://doi.org/10.1038/s41598-025-88871-y

Jocher, G., Chaurasia, A., & Qiu, J. (2023, January 7). *Ultralytics YOLO*. GitHub. https://github.com/ultralytics/ultralytics/blob/main/ultralytics/cfg/default.yaml

Lauro, M. D., Lina, Billy Marcelino, & Lorico Salim. (2024). Eksperimen Pemodelan dan Skenario Faster R-CNN untuk Penerapan Self Checkout Cashier. *JURNAL FASILKOM*, *14*(3), 715–721. https://doi.org/10.37859/jf.v14i3.7596

Li, Y., Xie, S., Chen, X., Dollar, P., He, K., & Girshick, R. (2021). Benchmarking detection transfer learning with vision transformers. *ArXiv Preprint ArXiv:2111.11429*. https://doi.org/10.48550/arXiv.2111.11429

Ramos, P., Ramos, R., & Garcia, N. (2025). Data leakage in visual datasets. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6309–6319.

Rane, N. (2023). YOLO and Faster R-CNN object detection for smart Industry 4.0 and Industry 5.0: applications, challenges, and opportunities. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4624206

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rumala, D. J. (2023). *How You Split Matters: Data Leakage and Subject Characteristics Studies in Longitudinal Brain MRI Analysis* (pp. 235–245). https://doi.org/10.1007/978-3-031-45249-9_23

Shailaja, K. (2020). Smart retail checkout. *International Journal of Research*, *7*(5).

Srivastava, M. M. (2020). Bag of Tricks for Retail Product Image Classification. In A. Campilho, F. Karray, & Z. Wang (Eds.), *Image Analysis and Recognition* (pp. 71–82). Springer International Publishing.

Srivastava, M. M., & Kumar, P. (2021). Machine Learning approaches to do size based reasoning on Retail Shelf objects to classify product variants. *ArXiv Preprint ArXiv:2110.03783*.

Tian, Y., Ye, Q., & Doermann, D. (2026). Yolov12: Attention-centric real-time object detectors. *Advances in Neural Information Processing Systems*, *38*, 78433–78457.

Wang, C., Huang, C., Zhu, X., & Zhao, L. (2022). One-Shot Retail Product Identification Based on Improved Siamese Neural Networks. *Circuits, Systems, and Signal Processing*, *41*(11), 6098–6112. https://doi.org/10.1007/s00034-022-02062-y

Wei, Y., Tran, S., Xu, S., Kang, B., & Springer, M. (2020). Deep Learning for Retail Product Recognition: Challenges and Techniques. *Computational Intelligence and Neuroscience*, *2020*, 1–23. https://doi.org/10.1155/2020/8875910

Wu, Y., Kirillov, A., Massa, F., Lo, W. Y., & Girshick, R. (2019). *Detectron2*. GitHub. https://github.com/facebookresearch/detectron2/blob/main/detectron2/config/defaults.py

Yazdanjouei, H., Mansouri, A., & Shokouhifar, M. (2025). A Co-Training Semi-Supervised Framework Using Faster R-CNN and YOLO Networks for Object Detection in Densely Packed Retail Images. *ArXiv Preprint ArXiv:2509.09750*. https://doi.org/10.48550/arXiv.2509.09750