An improved real time detection transformer method for retail product detection

Andi Wahyu Maulana[1,2], Suryo Adhi Wibowo[1,2]
[1]School of Electrical Engineering, Telkom University, Bandung, Indonesia
[2]Center of Excellence Artificial Intelligence for Learning and Optimization (CoE AILO), Telkom University, Bandung, Indonesia

Article Info

Article history:

Received Jul 9, 2024
Revised Jul 14, 2025
Accepted Aug 6, 2025

Keywords:

Detection transformer
Fusion block
Intra-class variation
Object detection
Retail product

ABSTRACT

The main problem in retail product detection is intra-class variation, as some products have similar but distinct characteristics. The primary goal of this study is to address the problem of object detection on intra-class variation in retail environments. As a result, a new approach for object detection of retail products was developed by modifying the real time detection transformer (RT-DETR) model. To manage intra-class variation more successfully, the RT-DETR model is updated by modifying its architecture. There are two convolutions in the contextual cross-feature module (CCFM) fusion block section, which is adjusted by adding one convolution layer to each CCFM fusion block. A customized dataset was meticulously constructed to reflect the wide range of products frequently seen in retail outlets. For the constructed datasets, tests were run using the mean average precision (mAP) metric, which had a mAP@0.5 of 99.5% and a mAP@0.5:0.95 of 88.2%. The updated model is superior compared to original model. The difference in mAP@0.5:0.95 was 2.5%, while precision increased by 1.3% and recall increased by 0.1%. Although the mAP0.5 results stay unchanged, the gains in the other metrics suggest that the RT-DETR model modifications can improve object detection skills, particularly when dealing with intra-class variation in retail merchandise.

Corresponding Author:

Suryo Adhi Wibowo
School of Electrical Engineering, Telkom University
Bandung, Indonesia
Email: suryoadhiwibowo@telkomuniversity.ac.id

1. INTRODUCTION

In the era of globalization and technological innovation, the retail industry has undergone significant transformation, driven by changing consumer behaviors and intensifying market competition. Deep learning approaches for product item detection have emerged as a critical technological solution [1], addressing the complex challenges of modern retail environments. Object detection, a fundamental branch of computer vision, aims to identify and localize specific objects within images or videos [2], becoming increasingly crucial for understanding consumer interactions, optimizing shopping experiences, and managing inventory.

The retail sector continuously evolves [3] to meet dynamic market demands, with technological efficiency emerging as a key differentiator. Smart cart technologies represent a promising frontier in this technological revolution, offering solutions to streamline the shopping experience by leveraging advanced object recognition capabilities [4]. These intelligent systems can automatically identify products, reduce checkout times, minimize pricing errors, and provide real-time product information, fundamentally transforming traditional retail interactions [5].

Object detection in smart carts provides a number of crucial benefits. First of all, it increases shopping speed by shortening the check-out process [6]. Second, it reduces the possibility of inaccuracies in registering products and pricing. Third, it allows businesses to immediately provide promotional material or product recommendations to consumers. In the field of retail shopping, the availability and accuracy of product information is crucial. When a customer uses a smart cart to shop, an object detection system recognizes and records the products that are inserted or taken out from the cart. While plenty of research has been done in the subject of item detection, specific problems occur in the setting of dynamic retail environments, where variations in product forms, colors, and groupings can be challenging.

In comparison with prior research, Santra *et al.*'s study [7] used a reconstruction-classification network (RC-Net) approach, which combines classification and reconstruction tasks to enhance classification accuracy. The reconstruction step focuses on reducing noise and enhancing image quality, while classification aims for precise recognition of objects. RC-Net has proven effective in handling image quality variations and enhancing overall classification performance. On the evaluated datasets, the method's accuracy rate was approximately 90%; however, it still requires improvement to continuously exceed 80% accuracy in all circumstances. Conversely, Hsia *et al.* [8] used data augmentation in conjunction with the faster region-based convolutional neural network (R-CNN) technique in their experiment. To increase the model's efficiency, they used techniques including rotation, flipping, and scaling to diversify the dataset. Their findings confirmed that these augmentation methods significantly boosted the model's accuracy and made it more robust against input variations. Because of the small quantity of the dataset, the model had trouble identifying very tiny product differences, even though it achieved a high mean average precision (mAP) accuracy of 99.27%. Lee *et al.* [9] stated, for retail product detection, they used the mobile neural network version 3 (MobileNet V3) architecture in conjunction with the you only look once version 5 (YOLOv5) model. They employed methods like rotation, flipping, and scaling to diversify the dataset, boosting the model's effectiveness. Findings demonstrated that data augmentation notably enhanced both accuracy and resilience of the model to diverse input variations. Despite achieving a high mAP accuracy of 99.27%, the model encountered challenges in distinguishing extremely subtle product differences, primarily due to the dataset's limited size. Lee *et al.* [9] utilized the YOLOv5 model combined with the MobileNet V3 architecture for retail product detection. The goal of this combination was to maximize detection efficiency and speed without sacrificing accuracy. According to experimental results, this model is perfect for real-time retail applications since it can reliably and swiftly recognize retail products. The study had constraints because of the relatively modest scale of the datasets employed, even though it achieved a 98.5% mAP accuracy.

Based on prior research findings, this study will present a modified real time detection transformer (RT-DETR) model to improve mAP [10] accuracy utilizing a self-processed dataset based on retail products in the Indonesian product business in real time. In addition to the self-generated dataset, the model will be tested with three other datasets: the grocery dataset [11], which focuses on products with different unit sizes; the retail product checkout (RPC)-dataset [12], which challenges products with similar characteristics; and the densely segmented supermarket (D2S)-dataset [13], which tests detection under different lighting conditions and product stacks. The proposed approach focuses on examining product variations across size, color, and type, with a primary objective of achieving a mAP accuracy exceeding 90%.

The key contributions of this study include: i) a novel architectural adaptation of the RT-DETR model [14], ii) development of a comprehensive six-class dataset representing Indonesian retail products, and iii) a robust methodology for detecting products with highly similar attributes across different categories. This research advances real-time product detection capabilities. It seeks to provide a sophisticated solution that can significantly enhance retail technology's precision and effectiveness.

2. METHOD

Retail product detection is used to identify and classify products from images. This process is useful for applications such as smart carts and inventory management systems. We propose a method based on a modified RT-DETR model. The main procedures in this research include data collection using a turntable setup and a Fujifilm X-T20 camera, data augmentation to enhance dataset diversity, and model training with customized RT-DETR layers to improve detection accuracy. The final output is a reliable product detection system. The main procedures are shown in Figure 1.

2.1. Data collection

The dataset for Indonesian retail products was collected using a Fujifilm X-T20 camera, with the imaging process facilitated by a turntable with paper on it, allowing for the rotation of products to capture images from multiple angles. This method ensures comprehensive visual data representation from various perspectives. The dataset consists of 6 product classes: Buavita Guava, Chitato Lite Seaweed, Oreo Original, Red Bull Drink, Chocolate Wheat Essence, and Selai Olai Strawberry, with each class having 65 images taken from different angles. The camera settings were optimized for product photography, including a medium aperture for depth of field and sharpness, an adjusted shutter speed to avoid motion blur, and a low international organization for standardization (ISO) to minimize noise. Proper lighting was ensured using diffused light sources to avoid harsh shadows and reflections, with brightness and exposure calibrated for natural colors and adequate contrast. The turntable allowed for systematic and controlled rotation, typically set to a fixed degree increment, ensuring consistent angles and intervals for each product. The images were composed to keep the product centered and at a consistent distance from the camera, with plain backgrounds to avoid distractions. This approach resulted in a high-quality, consistent dataset suitable for various applications in computer vision and deep learning. The results can be seen in Figure 2.

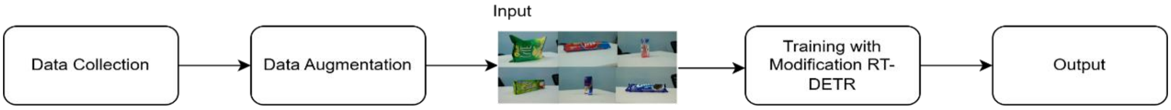


Figure 1. The proposed methodology

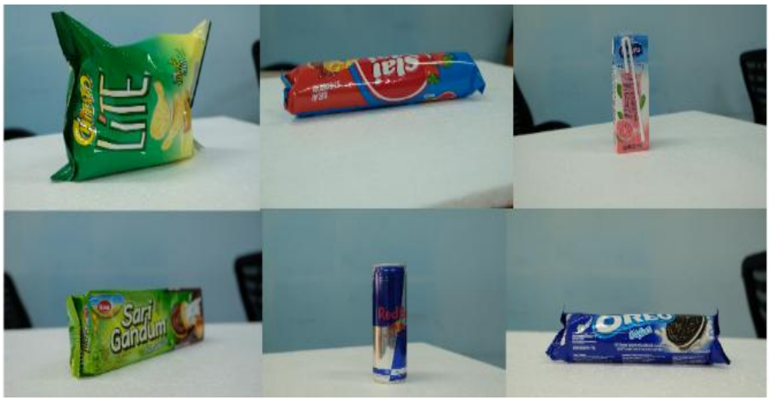


Figure 2. The results of the dataset image capture

2.2. Data augmentation

Image augmentation plays a crucial role in enhancing a dataset by providing a wider variety of examples. These examples help a model generalize better, especially when dealing with intra-class variations. By artificially enlarging the dataset, the model becomes more robust and capable of handling diverse scenarios, thus reducing the risk of overfitting and improving overall performance.

In this project, several augmentation techniques were applied to address common challenges encountered in retail product datasets, for the augmentation settings can be seen in Table 1. Auto orientation ensures that all images are properly aligned, which is essential for consistent training. Resizing the images to 640×640 pixels standardize dimensions, making the dataset uniform and reducing computational load. Flipping images horizontally and vertically introduces variations in product orientation, helping the model recognize items regardless of their placement. Cropping with a minimum zoom of 0% and a maximum zoom of 70% simulates different distances and perspectives, enhancing the model's ability to detect products at various zoom levels.

Table 1. Pre-processing and augmentation

Type	Process	Setup
Pre-processing	Auto oriented	Applied
Pre-processing	Resize	Stretch to 640×640
Augmentation	Flip	Horizontal, vertical
Augmentation	Crop	0% minimum zoom, 70% maximum zoom
Augmentation	Rotation	Between -45° and +45°
Augmentation	Shear	±15° horizontal, ±15° vertical
Augmentation	Brightness	Between -30% and +30%
Augmentation	Blur	Up to 2.5 px
Augmentation	Cutout	20 boxes with 10% size each

Rotation between -45° and +45° accounts for rotational differences, ensuring the model can identify products from various angles. Shearing both horizontally and vertically by ±15° distorts the image slightly, mimicking real-world distortions and improving robustness. Brightness adjustments between -30% and +30% allow the model to perform well under varying illumination conditions. Blur up to 2.5 pixels adds slight blurring to simulate motion or focus variations, making the model resilient to such issues. Finally, cutout with 20 boxes each sized at 10% covers random parts of the image to simulate occlusion, training the model to recognize products even if partially obscured.

The augmentation process was carried out using Roboflow [15], a platform that simplifies dataset modification. This comprehensive augmentation strategy resulted in a training dataset of 780 photos and a validation dataset of 75 images. Additionally, a testing dataset of 38 images was generated, totaling 893 images.

2.3. Public datasets

The study employs three public datasets to comprehensively validate the RT-DETR model's performance in addressing intra-class variation challenges. The grocery dataset [11], comprising 33,919 images with nearly identical product features, provides a rigorous test for detecting subtle product variations. The RPC-dataset [12], with its expansive 200 product classes and 83,699 images, offers a large-scale challenge in retail product detection, while the D2S-dataset [13], though smaller with 3,729 images, introduces complex detection scenarios through varied lighting conditions and product stacking. These datasets collectively represent a comprehensive evaluation framework, enabling a robust assessment of the model's capability to accurately recognize and distinguish products with highly similar characteristics across different contexts. The dataset partitioning follows a standard machine learning approach: the grocery dataset is split 85% for training, 10% for validation, and 5% for testing; the RPC-Dataset uses a 70/20/10 split; and the D2S-dataset maintains the same 70/20/10 distribution. This strategic selection and partitioning of datasets ensure a comprehensive validation of the proposed RT-DETR model, addressing key challenges in retail product detection such as intra-class variation, product similarity, and variations in image capture conditions.

2.4. Real-time detection transformer

The RT-DETR [14] is a real-time vision transformer (ViT) [16] model made up of three core components: a backbone, a hybrid encoder, and a decoder transformer that also includes an extra prediction head. Figure 3 shows the system's structure. This model uses the output features from the final three backbone stages (S3, S4, and S5) as input for the encoder [14]. Through intra-scale interaction [17] and inter-scale fusion [18], the hybrid encoder [19] converts multi-scale features [20] into a series of image-level features [21]. Then, an intersection of union (IoU)-aware query selection method [22] is applied to extract features from the encoder's output as the initial object query for the decoder [23]. The decoder then refines these queries step by step to produce bounding boxes and confidence scores. To boost both accuracy and efficiency, the model uses attention-based intrascale feature interaction (AIFI) [24] and CNN-based cross-scale feature fusion (CCFM) [25]. AIFI helps cut down redundancy at stage S5 while still capturing the relationships between high-level semantic features, which supports object detection. The model also skips low-level intra-scale interactions because they lack semantic meaning and can cause duplication issues [26]. RT-DETR also tackles inconsistencies between classification scores and IoU confidence distributions [27]. During training, the model is designed to link high IoU scores to high classification scores, which helps prevent inaccurate predictions and avoids selecting boxes that have low IoU scores even if they have high classification scores [14]. This optimization improves performance by aligning classification and location confidence effectively. The detector optimization goal can be rephrased in (1).

$$L(\hat{y}, y) = L_{box}(\hat{b}, b) + L_{cls}(\hat{c}, \hat{b}, y, b) = L_{box}(\hat{b}, b) + L_{cls}(\hat{c}, c, IoU) \quad (1)$$

Where $\hat{y}$ and y denote prediction and ground truth, $\hat{y}$={ c, $\hat{b}$} and y={$\hat{c}$, $\hat{b}$}, c and b represent categories and bounding boxes, respectively [14].

2.5. Real-time detection transformer modification

Figure 4 illustrates the fusion block employed in the CFFM framework, which is specifically designed to enhance feature interactions and improve overall model performance. As shown in Figure 4(a), there is a fusion block that aims to combine adjacent features into new features. This fusion block contains n repblocks [14] and the outputs of two paths are fused through sequential addition of elements. The improved fusion block is depicted in Figure 4(b), each fusion block gains one more convolution layer to improve image object detection accuracy. Convolution layers are used in image processing to recognize local patterns and allow the model to understand increasingly complicated feature hierarchies. The convolution layer can extract significant image elements including edges, textures, and other visual patterns [28]. This assists the model in comprehending the distinct qualities of the object being recognized. This improvement is likely to aid in object detection, particularly for things with comparable but distinct variations.

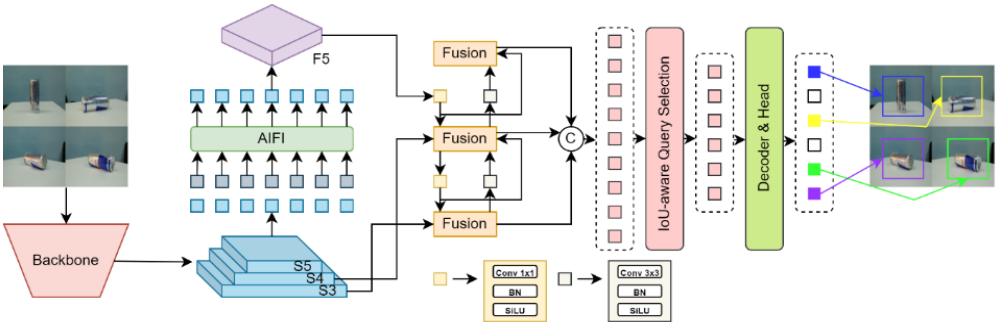
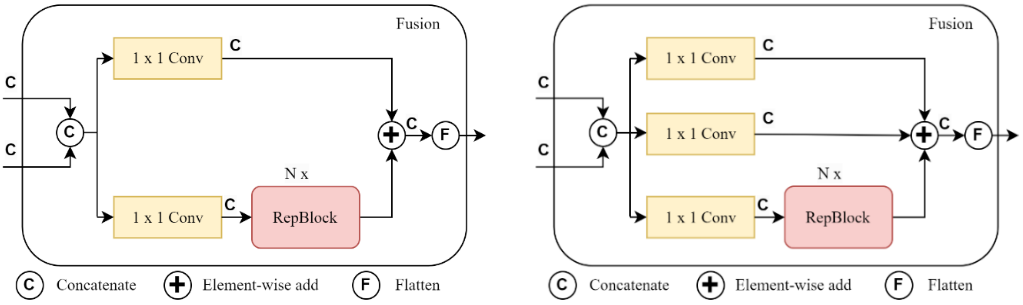


Figure 3. The RT-DETR architecture



(a) (b)

Figure 4. The fusion block in the CCFM, which is designed to enhance feature interaction for improved model performance consists of two components: (a) original fusion block and (b) modification fusion block

2.6. Performance parameter

Precision, recall, and mAP are some frequently used metrics in assessing the effectiveness of machine learning models, especially in the context of object detection. These parameters help understand how well the model detects and recognizes the desired objects. The ratio of true positives, or accurate forecasts of real objects, to all positive predictions, including inaccurate ones, is known as precision. Precision gauges how accurate the model is at making predictions; that is, what proportion of all the model's positive predictions are true in (2).

$$Precession = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \tag{2}$$

Recall is the ratio of the number of true positives to the total number of actual objects (the sum of true positives and false negatives). Recall measures the model’s ability to find all instances of the object, as depicted in (3).

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{3}$$

The average of the average precision (AP) over all classes is known as mean average precision, or mAP. For every class, AP is the area under the precision-recall curve. To give a thorough picture of the model's effectiveness in identifying objects from all tested classes, mAP integrates accuracy and recall. AP is calculated for each class and then averaged as part of the computation procedure. The mAP formula in (4).

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

Where N is the number of classes and AP_i is the AP for class i.

3. RESULTS AND DISCUSSION

3.1. Experiment preparation

Before training, numerous settings in the RT-DETR model configuration section must be made. These parameters will then be checked on all datasets. Table 2 displays the parameter settings. The NVIDIA DGX A100 device is used for training in this test.

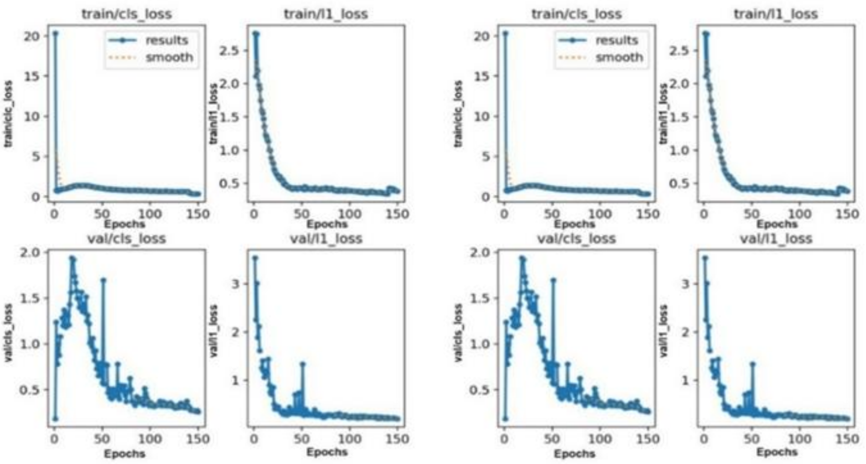
Table 2. Training parameter setup

Parameters	Value
Epochs	150
Batch size	16
Optimizer	Auto
Initial learning rate	1×10^{-2}

3.2. Our dataset result and discussion

This study addresses gaps identified in previous research, such as those by Santra *et al.* [7], Hsia *et al.* [8], and Lee *et al.* [9], who utilized various architectures like RC-Net, Faster R-CNN, and YOLOv5 for retail product detection. While these methods improved object detection accuracy, they did not fully address challenges related to intra-class variation, such as distinguishing subtle differences between similar products. This research focuses on modifying the RT-DETR model to better handle these challenges, particularly in enhancing object detection accuracy by accounting for intra-class variations.

The dataset was trained on the modified RT-DETR model for 150 epochs. Figure 5 presents the performance evaluation graphs of the proposed model, illustrating its effectiveness across different metrics. Specially, Figure 5(a) illustrate classification loss and L1 loss. Classification loss estimates the correct object category in the bounding box, while L1 loss calculates the absolute difference between expected and target values. The loss results for classification and L1 are stable during training but unstable during validation, stabilizing after 100 epochs. The dataset used is self-generated, and augmentation caused some instability, particularly in generalizing data unseen during training. Accuracy results of the updated RT-DETR model, including precision, recall, mAP@0.5, and mAP@0.5:0.95, are shown in Figure 5(b). These results are good, but learning iteration stability improves after 100 epochs due to the dataset's cutout, allowing better recognition of products and handling of intra-class variation. Table 3 displays the results of training three models: YOLOv8, RT-DETR, and the modified RT-DETR. The results show that when the dataset is run with the YOLOv8 model, precision accuracy reaches 91.1%, recall is approximately 94.07%, mAP@0.5 is high at 98.7%, but mAP@0.5:0.95 is relatively low at 77.4%. The RT-DETR model outperforms YOLOv8 with a precision of 97.4%, recall of 99.5%, mAP@0.5 of 99.5%, and mAP@0.5:0.95 of 85.7%. The highest performance is achieved by the modified RT-DETR model, which has a precision of 98.7%, recall of 99.6% (a 0.1% improvement over RT-DETR), mAP@0.5 of 99.5% (matching the original RT-DETR), and mAP@0.5:0.95 of 88.2%, surpassing the original RT-DETR model.



(a) (b)

Figure 5. Performance evaluation graphs of the proposed model: (a) classification and L1 loss graph for train and validation and (b) accuracy graph of precision, recall, mAP@0.5 and mAP@0.5:0.95

Table 3. Result and comparison in our dataset

Method	Precission (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8 [29]	91.1	94.07	98.7	77.4
RT-DETR [29]	97.4	99.5	99.5	85.7
RT-DETR Mod	98.7	99.6	99.5	88.2

The modified RT-DETR model proves that it can outperform the original model in the field of object detection that pays attention to the intra-class variation part. Figure 6 shows the validation results of the modified RT-DETR model. The validation results are shown in Figure 6(a). There are six class objects found, and the predictions are all correct. The comparison findings of each class from the dataset are displayed in Table 4. In terms of mAP@0.5 accuracy, it appears to be 99.5% across all classes. A clearer comparison can be seen in the accuracy of mAP@0.5:0.95, in which each class is exceeded by the modified RT-DETR model. It is apparent that the improved RT-DETR model produces better results than the original model, and this modification has proven that the effect of adding convolution layers to the CCFM fusion block can increase model performance accuracy.

Table 4. Comparison of the results for each class for the RT-DETR and modified RT-DETR models

Class	RT-DETR		RT-DETR modified	
	mAP@0.5 (%)	mAP@0.5:0.95 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Buavita Jambu	99.5	74.4	99.5	79.4
Chitato Lite Rumput Laut	99.5	90.3	99.5	92.4
Oreo Original	99.5	88.3	99.5	91.3
Red Bull Drink	99.5	88.2	99.5	90.6
Sari Gandum Cokelat	99.5	86.1	99.5	86.6
Slai Olai Stroberi	99.5	86.9	99.5	89.1

3.3. Other dataset result and discussion

3.3.1. Grocery dataset result and discussion

Table 5 presents the results of training YOLOv8, RT-DETR and modified RT-DETR models. The dataset is run with the YOLOv8 model the result of precision accuracy is 99.8%, the result of recall reaches around 99.8%, mAP@0.5 is 99.4% and for mAP@0.5:0.95 is 82.1%. Then the RT-DETR model produces a precision value of 99.8% the same as the YOLOv8 model, the recall value reaches 99.9%, the mAP@0.5 value reaches 99.5% and mAP@0.5:0.95 reaches 83.7% indicating that the RT-DETR model is superior to the YOLOv8 model. The highest performance is achieved by the modified RT-DETR model, which has a precision of 99.9% (a 0.1% improvement over the original RT-DETR), recall of 99.9% (matching the original RT-DETR), mAP@0.5 of 99.5% (the same as the original RT-DETR), and mAP@0.5:0.95 of 84.2%, surpassing the original model. The modified RT-DETR model proves that it can outperform the original model in the field of detection objects that pay attention to the intra-class variation part. For the validation results can be seen in Figure 6(b).

3.3.2. Retail product checkout dataset result and discussion

Table 6 presents the results of training YOLOv8, RT-DETR and modified RT-DETR models. The dataset is run with the YOLOv8 model the result of precision accuracy is 99.8%, the result of recall reaches around 99.8%, mAP@0.5 is 99.2% and for mAP@0.5:0.95 is 86.4%. Then the RT-DETR model produces a precision value of 99.8% the same as the YOLOV8 model, the recall value reaches 99.8%, the mAP@0.5 value reaches 99.4% and mAP@0.5:0.95 reaches 88.03% with this stating that the RT-DETR model is superior to the YOLOv8 model. The highest performance is achieved by the modified RT-DETR model, which has a precision of 99.9% (a 0.1% improvement over both the original RT-DETR and YOLOv8 models), recall of 99.8% (the same as the original RT-DETR and YOLOv8), mAP@0.5 of 99.5%, and mAP@0.5:0.95 of 88.2%, surpassing the original RT-DETR model. The modified RT-DETR model proves that it can outperform the original model in the field of detection objects that pay attention to the intra-class variation part. For the validation results can be seen in Figure 6(c).

Table 5. Result and comparison in Grocery dataset

Method	Precission (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8 [29]	99.8	99.8	99.4	82.1
RT-DETR [29]	99.8	99.9	99.5	83.7
RT-DETR Mod	99.9	99.9	99.5	84.2

Table 6. Result and comparison in RPC dataset

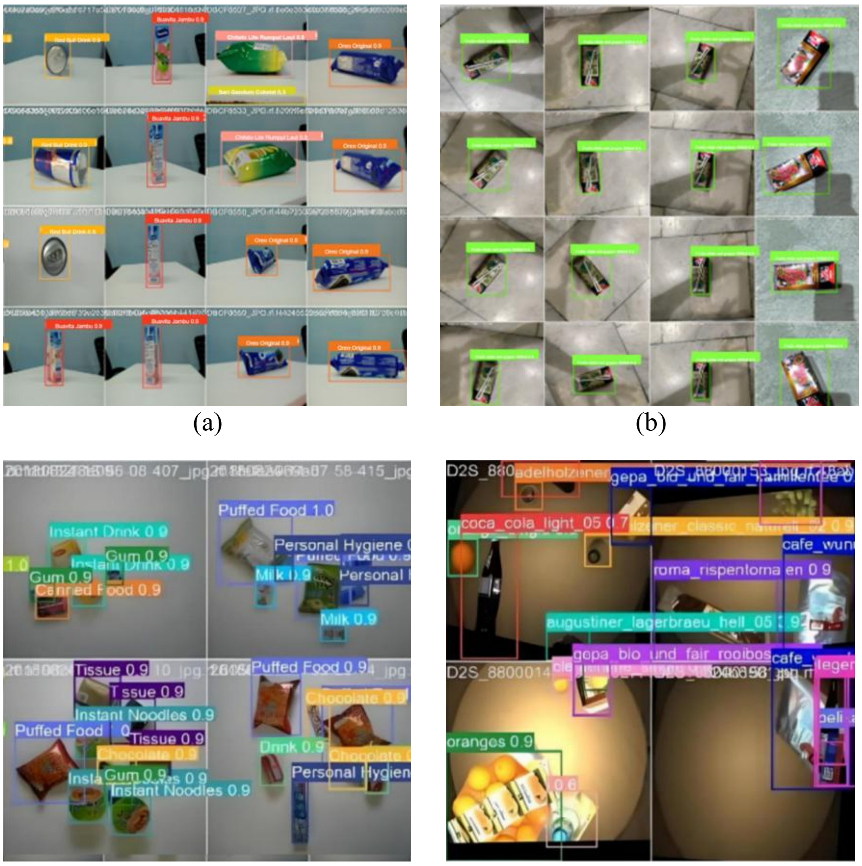
Method	Precission (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8 [29]	99.8	99.8	99.2	86.4
RT-DETR [29]	99.8	99.8	99.4	88.03
RT-DETR Mod	99.9	99.8	99.5	88.2

3.3.3. Densely segmented supermarket dataset result and discussion

Table 7 presents the results of training YOLOv8, RT-DETR and modified RT-DETR models. The dataset is run with the YOLOv8 model the result of precision accuracy is 91.8%, the result of recall reaches around 90.9%, mAP@0.5 is 81.9% and for mAP@0.5:0.95 is 58.2%. Then the RT-DETR model produces a precision value of 93.9%, the recall value reaches 84.05%, the mAP@0.5 value reaches 91.8% and mAP@0.5:0.95 reaches 72.03% with this stating that the RT-DETR model is superior to the YOLOv8 model. The highest performance is achieved by the modified RT-DETR model, which has a precision of 94.1% (0.2% higher than the original RT-DETR), recall of 85.5%, mAP@0.5 of 92.2%, and mAP@0.5:0.95 of 70.6%. Although the mAP@0.5:0.95 is slightly lower than the original model, the modified RT-DETR outperforms the original RT-DETR in precision, recall, and mAP@0.5, making it the superior model overall. The slightly lower accuracy in this dataset compared to previous tests is due to the dataset containing products with variations in lighting, which affected the training process. The modified RT-DETR model proves that it can outperform the original model in the field of detection objects that pay attention to the intra-class variation part. The validation results can be seen in Figure 6(d).

Table 7. Result and comparison in D2S dataset

Method	Precission (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8 [29]	91.8	90.9	81.9	58.2
RT-DETR [29]	93.9	84.05	91.8	72.03
RT-DETR Mod	94.1	85.5	92.2	70.6



(a) (b)

(c) (d)

Figure 6. Validation results of the RT-DETR modification model on (a) our dataset, (b) Grocery dataset, (c) RPC-dataset, and (d) D2S-dataset

4. CONCLUSION

Based on the results of this study, it is possible to conclude that the RT-DETR model can be adjusted to increase the model's performance accuracy. We added a convolution layer to each fusion block in the CCFM fusion block to improve the accuracy of image object detection. Convolution layers are used in image processing to recognize local patterns and allow the model to understand increasingly complicated feature hierarchies. Convolution layers can extract significant information from images like as edges, textures, and other visual patterns, assisting the model in understanding the distinct qualities of the item being recognized. Testing our own dataset as well as three other datasets demonstrated that our modified RT-DETR model may increase the accuracy of product object detection and aid in the detection of product variations. The improved results of mAP@0.5 for the self-provided dataset reached 99.5% and mAP@0.5:0.95 reached 88.2% as a result of our modified RT-DETR model, which also applies to the other three types of datasets that outperformed the original model and YOLOv8. However, there are limitations to this study. The improvements observed in the modified RT-DETR model may not be consistent across all types of datasets, particularly those with more complex intra-class variations or extreme lighting conditions. Furthermore, the model's performance could be constrained by the size and diversity of the dataset used for training. For future work, we suggest exploring additional modifications to the RT-DETR architecture beyond the fusion block, such as incorporating advanced attention mechanisms or experimenting with other types of convolutional layers. Additionally, expanding the dataset with more diverse product categories and challenging environments could help further enhance the model's robustness and accuracy. Investigating the impact of different augmentation strategies and optimizing the training process could also lead to better generalization across various retail scenarios.

ACKNOWLEDGEMENTS

This work was supported by Telkom University.

FUNDING INFORMATION

This work was funded by the RIIM Batch 3 program grant from LPDP, Ministry of Finance, Republic of Indonesia, and BRIN (No. 81/IV/KS/05/2023).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Andi Wahyu Maulana	✓	✓	✓	✓	✓	✓	✓		✓		✓			
Suryo Adhi Wibowo	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

Not applicable.

DATA AVAILABILITY

The data that support the findings of this study are openly available in:

- Roboflow at https://universe.roboflow.com/new-workspace-wfzw3/grocery-dataset-q9fj2
- RPC Dataset Github at https://rpc-dataset.github.io/
- Mvtec Software at https://www.mvtec.com/company/research/datasets/mvtec-d2s

REFERENCES

[1] J. Thøgersen, "Consumer behavior and climate change: consumers need considerable assistance," *Current Opinion in Behavioral Sciences*, vol. 42, pp. 9–14, Dec. 2021, doi: 10.1016/j.cobeha.2021.02.008.

[2] Y. Amit, P. Felzenszwalb, and R. Girshick, "Object detection," in *Computer Vision*, vol. 5, no. 1, Cham: Springer International Publishing, 2020, pp. 1–9, doi: 10.1007/978-3-030-03243-2_660-1.

[3] W. M. Lim, S. Kumar, N. Pandey, D. Verma, and D. Kumar, "Evolution and trends in consumer behaviour: insights from journal of consumer behaviour," *Journal of Consumer Behaviour*, vol. 22, no. 1, pp. 217–232, Jan. 2023, doi: 10.1002/cb.2118.

[4] D. Grewal, S. Benoit, S. M. Noble, A. Guha, C. P. Ahlbom, and J. Nordfält, "Leveraging in-store technology and AI: increasing customer and employee efficiency and enhancing their experiences," *Journal of Retailing*, vol. 99, no. 4, pp. 487–504, Dec. 2023, doi: 10.1016/j.jretai.2023.10.002.

[5] H. B.-Salau, A. J. Onumanyi, D. Michael, R. Isa, C. O. Alenoghena, and H. Ohize, "A new automated smart cart system for modern shopping centres," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2028–2036, Aug. 2021, doi: 10.11591/EEI.V10I4.2762.

[6] N. X. Jie and I. F. B. Kamsin, "Self- checkout service with RFID technology in supermarket," in *Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, 2021, vol. 4, doi: 10.2991/ahis.k.210913.062.

[7] B. Santra, A. K. Shaw, and D. P. Mukherjee, "Part-based annotation-free fine-grained classification of images of retail products," *Pattern Recognition*, vol. 121, Jan. 2022, doi: 10.1016/j.patcog.2021.108257.

[8] C. H. Hsia, T. H. W. Chang, C. Y. Chiang, and H. T. Chan, "Mask R-CNN with new data augmentation features for smart detection of retail products," *Applied Sciences*, vol. 12, no. 6, Mar. 2022, doi: 10.3390/app12062902.

[9] R. Y. Lee, S. Y. Chua, Y. L. Lai, T. Y. Chai, S. Y. Wai, and S. C. Haw, "Cashierless checkout vision system for smart retail using deep learning," *Journal of System and Management Sciences*, vol. 12, no. 4, pp. 232–250, Aug. 2022, doi: 10.33168/JSMS.2022.0415.

[10] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[11] new-workspace-wfzw3, "Grocery dataset computer vision model," *Roboflow Universe*, 2022. Accessed: Feb. 7, 2024. [Online]. Available: https://universe.roboflow.com/new-workspace-wfzw3/grocery-dataset-q9fj2

[12] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: a large-scale retail product checkout dataset," *arXiv-Computer Science*, 2019, [Online]. Available: http://arxiv.org/abs/1901.07249

[13] P. Follmann, T. Böttger, P. Härtinger, R. König, and M. Ulrich, "MVTec D2S: densely segmented supermarket dataset," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11214 LNCS, 2018, pp. 581–597, doi: 10.1007/978-3-030-01249-6_35.

[14] Y. Zhao *et al.*, "DETRs beat YOLOs on real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024, doi: 10.1109/CVPR52733.2024.01605.

[15] Q. Lin, G. Ye, J. Wang, and H. Liu, "RoboFlow: a data-centric workflow management system for developing AI-enhanced robots," *Proceedings of Machine Learning Research*, vol. 164, pp. 1789–1794, 2021.

[16] K. Islam, "Recent advances in vision transformer: a survey and outlook of recent work," *arXiv-Computer Science*, 2023. [Online]. Available: https://arxiv.org/abs/2203.01536

[17] J. Lin, X. Mao, Y. Chen, L. Xu, Y. He, and H. Xue, "D^2ETR: decoder-only DETR with computationally efficient cross-scale attention," *arXiv-Computer Science*, 2022. [Online]. Available: https://arxiv.org/abs/2203.00860

[18] C. Wang, X. Xing, Y. Wu, Z. Su, and J. Chen, "DCSFN: deep cross-scale fusion network for single image rain removal," in *MM 2020-Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, pp. 1643–1651, doi: 10.1145/3394171.3413820.

[19] L. An, L. Wang, and Y. Li, "HEA-Net: attention and MLP hybrid encoder architecture for medical image segmentation," *Sensors*, vol. 22, no. 18, Sep. 2022, doi: 10.3390/s22187024.

[20] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1369–1380, 2020, doi: 10.1007/s10044-019-00845-9.

[21] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6877–6886, 2021, doi: 10.1109/CVPR46437.2021.00681.

[22] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019, doi: 10.1109/CVPR.2019.00075.

[23] Z. Yao, J. Ai, B. Li, and C. Zhang, "Efficient DETR: improving end-to-end object detector with dense prior," *arXiv-Computer Science*, 2021. https://arxiv.org/abs/2104.01318.

[24] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2022.3168331.

[25] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, "CF2PN: a cross-scale feature fusion pyramid network based remote sensing target detection," *Remote Sensing*, vol. 13, no. 5, pp. 1–23, Feb. 2021, doi: 10.3390/rs13050847.

[26] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6163–6171, 2018, doi: 10.1109/CVPR.2018.00645.

[27] F. Kuppers, J. Kronenberger, A. Shantia, and A. Haselhoff, “Multivariate confidence calibration for object detection,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1322–1330, 2020, doi: 10.1109/CVPRW50498.2020.00171.

[28] J. Cao *et al.*, “DO-Conv: depthwise over-parameterized convolutional layer,” *IEEE Transactions on Image Processing*, vol. 31, pp. 3726–3736, 2022, doi: 10.1109/TIP.2022.3175432.

[29] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” *Ultralytics Inc.*, 2023. Accessed: Feb. 7, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolov8/

BIOGRAPHIES OF AUTHORS

Andi Wahyu Maulana received the M.T. degree in telecommunication engineering from Telkom University, Indonesia, in 2024. His research interests include machine learning, computer vision, data engineering, data analyst, data scientist, artificial intelligence. He can be contacted at email: awahyumaulana@student.telkomuniversity.ac.id.

Suryo Adhi Wibowo received a Ph.D. degree from the Department of Electrical and Computer Engineering, Pusan National University, Rep. of Korea, in 2018. He received the best student presentation award from the joint 8th International Conference on Soft Computing and Intelligent Systems and the 17th International Symposium on Advanced Intelligent Systems, Hokkaido, Japan, in 2016. His research interests are computer vision, computer graphics, pattern recognition, virtual reality, and machine learning. Mr. Suryo is a member of IEEE Signal Processing Society and a member of IEEE. He can be contacted at email: suryoadhiwibowo@telkomuniversity.ac.id.