Counter Vision: Self-Checkout System

Siddhesh Darak[*1], Ninad Chobe[#2], Dhruv Mehta[#3], Divanshu Uppal[#4], Sachin Sonawane[$5]
[*]*Student, Department of Computer Engineering, NMIMS University*
[#]*Student, Department of Artificial Intelligence and Machine Learning, NMIMS University*
[$]*Assistant Professor, Department of Artificial Intelligence and Machine Learning, NMIMS University*

Abstract*—Tremendous change has happened in the retail industry, especially in higher demand for efficient, touchless, and friendly systems of checkout [4]. This work proposes a camera-based self-checkout system known as CounterVision. This must replace common methods of barcodes by using advanced computer vision and deep learning techniques [8]. CounterVision employs the object detection model called YOLOv7 to instantly identify and process several products in challenging retail scenarios that depend on the changing lighting conditions and filled shelves [1]. In this work, the system is used with a mounted single camera with background subtraction and bounding box tracking to minimize the computational costs with high discovery accuracy [8]. Extensive testing resulted in the system achieving a mean Average Precision of 1.0 and recall of 0.97 at 0.5 IoU and 0.4 confidence threshold, underlining its robustness and reliability [1][8]. Although it does away with bar codes, CounterVision is one of the automated retailing solutions for cost-effectiveness by eliminating waiting time and improving the speed of operations; this means human error is eliminated [8]. This paper explores the actualities of scalability and diversity in datasets and more potential upgrades and use cases that can include edge computing and real-time connectivity for inventory management systems [5][8]. Reconfiguring checkout enables CounterVision to make retail places smarter, faster, and more reliable [4].

Keywords*—Self-checkout systems, Retail automation, Object detection, YOLOv7, Computer vision, Deep learning, Product recognition, Camera-based systems, Real-time processing, Non-barcode checkout, Retail efficiency, Contactless checkout, Smart retail technology, Dataset diversity, Inventory management, AI-driven retail solutions.

I. INTRODUCTION

A. Background

The retail business has become more dynamic, and today it requires establishing systems that have optimized the customer experience in the aspects of operational efficiency and scalability [8]. Check-out systems that function based on the use of a manual barcode scanner have been in trend for quite some time. However, such systems are facing critical limitations. They are labor-intensive, prone to errors, and cause bottlenecks, especially during peak shopping hours [11]. The higher the consumer's expectation to have a perfect shopping experience, the more retailers are challenged against taking innovative steps to make this whole process cost-effective [5].

Among the basic contemporary retail inventions used to abolish reliance on human intervention is self-checkout systems. However, most of these systems rely on barcodes, RFID tags, or QR codes for identification [8]. The conventional self-checkout lanes demand physical contact between customers and will, hence, cause inconvenience [4]. The breakthroughs in computer vision and artificial intelligence open great opportunities for the much-needed revolution of retail operations with fully automated, camera-based solutions [1][8].

A camera-based system has therefore redefined the scope of possible applications for autonomous checkout technology: only with computer vision and deep learning models, such a system can recognize products visually—not through barcodes or other markers on items [4]. It also accelerates the check-out process and enriches the overall customer experience [11].

B. Problem Statement

Despite growing interest in automated retail technologies, this has remained a significant barrier to the broader deployment of solutions that already exist [5][11]. These barcode-based systems, so ubiquitously deployed, are virtually independent of the integrity of the barcode labels—thus they remain susceptible to possible damage, loss, and even destruction [4]. Solutions based on RFID technology are extremely expensive to deploy and maintain, making them impractical for more modest retailers [8].

However, at this time, solutions coming from AI are generally very resource-intensive, comprising many sensors, cameras, and computational components; therefore, it can spike the cost of deployment very sharply [5]. The technical requirements of real-time processing of product data also across dynamic retail environments introduce variability in lighting

conditions, occlusions from overlapping items, and a variety of retail item appearances [1][8].

Given these limitations, there is a pressing need for a self-checkout solution that combines the accuracy and scalability of AI with the cost-effectiveness and simplicity required for practical retail deployment [11]. Such a system must be capable of detecting and identifying products without relying on barcodes or additional hardware, ensuring adaptability across diverse retail environments [8].

C. Motivation

Such research motivation is based on an increasing demand for scalable, reliable, and cost-effective self-checkout systems [4][8]. Cost-cutting and operational efficiency are now held dear by retailers across the globe, but faster checkouts are becoming more in demand by consumers [11]. Camera-based self-checkout systems, powered by cutting-edge computer vision technologies, have the potential to address these dual needs [5][1].

The proposed system, CounterVision, utilizes advances in object detection to offer a complete and bar-code-free check-out experience [1]. Even with just one fixed camera, YOLOv7—the system presents itself in utmost accuracy with the diversity of conditions that it may operate and can thrive powerfully using varied products [8][1]. Scalability and cost-effectiveness make the system especially suited for small and medium-sized retailing businesses [11].

D. Scope of the Research

The scope of this research extends beyond the development of a single solution [4][8]. It aims to establish a framework for integrating AI-based self-checkout systems into diverse retail environments [5]. The study explores critical factors influencing the performance and adaptability of such systems, including:

- Dataset Development: An extensive dataset of product appearances, shapes, and size for training and validation of the object detection model [9].
- Algorithm Optimization: Fine-tune the YOLOv7 model and related algorithms to achieve the best performance as there are dynamics of light and occlusions among other things [1][5].
- System Evaluation: Thorough testing to determine the system's accuracy, speed, and reliability under controlled conditions as well as real-life conditions [11][1].
- Future Upgrade: Add into and integrate other edge computing capabilities to further enhance the real-time processing capability and expand the dataset to input non-standard retail products [5][8].

II. LITERATURE REVIEW

With a growing demand for automated solutions in retail, tremendous development has been witnessed in self-checkout (SCO) systems. These machines were designed to make checkouts more efficient and have evolved from traditional barcode-based techniques to modern image recognition methods, significantly improving their applicability and effectiveness in retail stores [5][11].

The early approaches to SCOs relied heavily on barcode-based automation, limiting their scope within the automation domain [4]. Singh et al. (2018) demonstrated that bounding box annotations could be integrated with YOLO-based object detection to achieve early-stage improvements in inventory identification with minimal human intervention [1]. More recently, Ellis (2023) introduced ARC-AI, utilizing YOLOv8 to enhance accuracy, ensure privacy, and provide scalable frameworks for dynamic retail environments [11].

A prominent trend in modern SCO systems is the use of image-based solutions powered by deep learning. Among the models promising both accuracy and efficiency is YOLOv7, which is particularly suited for real-time applications like retail checkout [8]. Wang et al. (2022) highlighted YOLOv7's ability to avoid overfitting using a "trainable bag-of-freebies," achieving state-of-the-art performance. This approach has been successfully applied in smart vending systems that rely on item recognition without barcodes, addressing issues like non-standard product appearances, as shown by Xia et al. (2021) [5].

Deep architectures such as MobileNet and YOLO have become crucial in SCO systems, especially in resource-constrained environments. Machavaram et al. (2024) and Tan et al. (2022) demonstrated that lightweight convolutional neural networks (CNNs) could deliver high-speed and accurate recognition systems even for complex real-world challenges [1][8]. SCOs face several challenges, including variations in lighting, occlusions, and diverse visual appearances of products. Robustness enhancements, such as mosaic augmentation and high-quality data labeling, have been developed to address these issues [4].

Thanks to curated datasets and optimized hyperparameters, YOLOv7-based systems have shown significantly lower detection errors under

controlled experimental conditions [5]. CounterVision, for example, combines high-performance detection algorithms with a comprehensive training regime, achieving exceptional precision. Its one-camera configuration utilizes YOLOv7 to operate without manual input of images or context, delivering a mean Average Precision (mAP) of 1.0 and a recall of 0.97, even under challenging conditions [11].

The research trend in SCOs centers around complex, image-based systems leveraging deep learning to achieve robust and efficient performance. Innovations like these are poised to revolutionize retail operations by offering customers seamless automated checkouts while addressing critical operational challenges for retailers. Future work aims to enhance scalability and adaptability for various retail environments, further transforming the retail landscape [8][1].

III. METHODOLOGY AND IMPLEMENTATION

The technical workflow for developing the YOLOv7 object detection system in your project is detailed and involves the crucial phases of Training and Predicting, each underpinned by specific methodologies for robust performance.

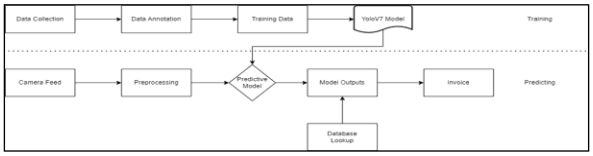

Fig. 1 Block Diagram of the suggested system

A. Training Phase

1) Data Collection: The project underlines a need for a diverse set of data so that the model will be able to generalize well over various scenarios and perspectives. For this paper, four one-minute videos came with 14 different products. They had been chosen with a notion to vary in shapes, sizes, and looks so that it could accommodate as many real-world conditions most commonly seen in retail.

2) Data Annotation We used labelling tools such as LabelMe, which involved writing bounding boxes on individual objects belonging to each frame of the video and then labelling them according to the class labels. The annotations in this structured JSON format are necessary for both spatial and categorical data to enable their usage during supervised training. With the annotation detailing object positions, high-quality input is provided to the model YOLOv7 during training, which enhances the detection and localization accuracy of the objects.

3) Training Configuration: The following configurations were followed for ensuring a generalized and efficient training.

- Learning Rate: Initiated at 0.001 for balanced updates during optimization.

- Momentum: Set to 0.937 to stabilize the weight updates and speed up convergence.

- Weight Decay: Fix 0.0005 is a regularization of model complexity to avoid overfitting.

- Warmup Epochs: Within each iteration, three epochs to stabilize the early learning stages.

- Data Augmentation: Mosaic Augmentation with a probability of 0.5, as sample image patches were merged to introduce feature diversity.

- Epochs and Batch Size-Incremented 100 epochs with a batch size of 64 to make the model as stable as possible.

It was trained, with the metrics mean Average Precision (mAP) and recall being monitored during the training to validate its effectiveness.

B. Training Phase:

1) Input and Preprocessing: The system acquires live and recorded inputs through an integrated camera. Pre-processing stages are resized and normalized before reaching the input of YOLOv7, enhancing the robustness of detection and standardizing the quality of inputs.

2) Inference: The trained YOLOv7 model processed all the video frames and detected the objects encircled by bounding boxes, labelled, and assigned with confidence scores. Overall, the system outperformed the reference system in Mean Average Precision (mAP) of 1.0 and a recall of 0.97 at the threshold of 0.5 IoU, which reflects the ability to handle complicated retail scenarios with changing lighting and occlusions.

3) Output & Integration: Objects detected are further analysed to provide database queries for metadata including product description and prices. These outputs are seamlessly integrated into automated invoice generation systems, enabling accurate billing with minimal human intervention.

C. Flowchart:

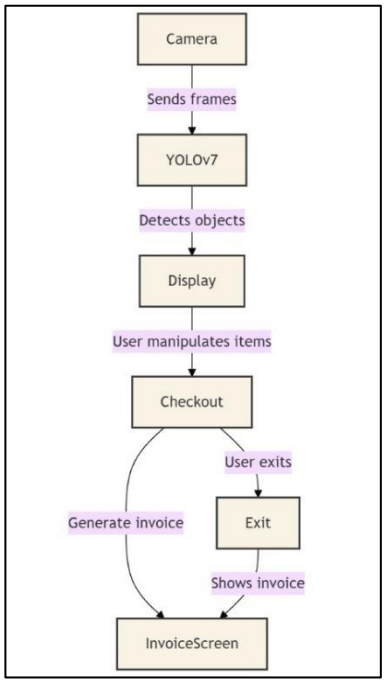

Fig. 2 Flowchart of the suggested system

The flowchart shows how the system works step by step to make the checkout process easy and efficient. It starts with a camera capturing video frames, which are sent to the YOLOv7 model to detect items in real time. The detected items are then shown on a display, where users can review and adjust quantities as needed. Once everything is finalized, the system moves to the checkout stage. Here, an invoice is generated, which the user can view on the screen. If the user chooses to exit, the system automatically shows the invoice. This process ensures that item detection, user adjustments, and payment are all handled smoothly, creating a fast and user-friendly experience.

IV. RESULT AND ANALYSIS

A. *Training Result:*

1) Confusion Matrix: It is a table that describes for each class of the dataset how many are true positives, false positives, false negatives, and true negatives. It visually represents the model's classification performance, highlighting areas where the model confuses one class for another. Confusion matrices in YOLOv7, although mainly developed for object detection, it brings out certain classes which require more training data or targeted fine-tuning. High values on the diagonal are correct classifications, whereas off-diagonal values indicate misclassifications.

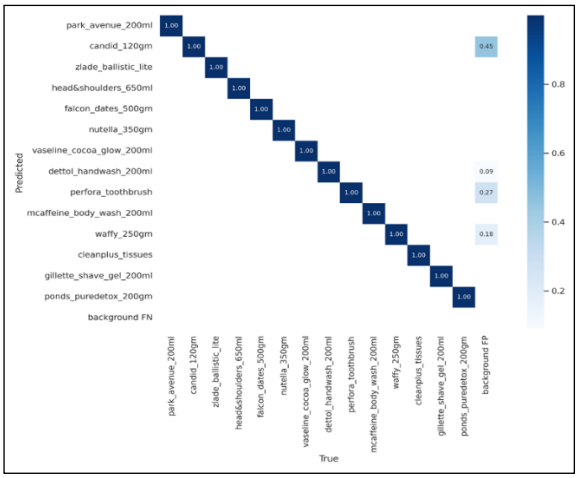

Fig. 3 Confusion Matrix

2) Precision Curve: (or P-Curve) shows where the model is precise at a certain confidence threshold. Precision calculates the number of true positives against all detections, both true positives and false positives. In object detection, high precision means that many of its predictions are correct with a very low number of false positives. A high, flat P Curve indicates that a model will achieve a robust trade-off between detection of true positives and minimization of false detections as the confidence threshold increases.

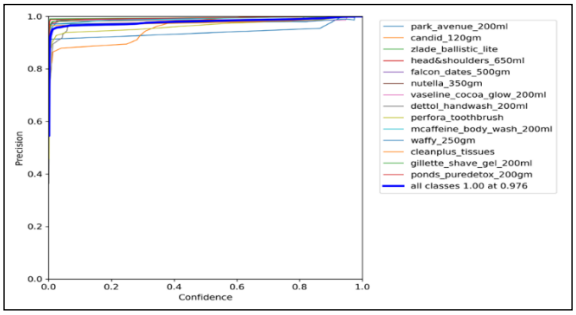

Fig. 4 Precision Curve

3) *Recall Curve: (or R-Curve)* is the graphical representation of how well the model classifies objects of interest correctly as a function of the confidence threshold considered. Recall is the fraction of actual positives that are correctly classified as such by the model; it depicts how well a model can locate all instances of relevance. A high recall on object detection means that the model recognizes most of the true objects within the images, though it may also include some false detections. A flat high R Curve means that the model always identifies the correct counts of objects at all the levels of shifted confidence thresholds.

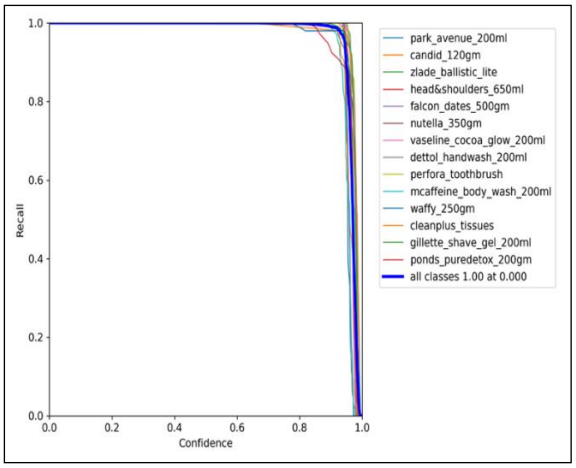

Fig. 5 Recall Curve

4) Precision-Recall Curve: (or PR-Curve) is a composite representation of both precision and recall across varying confidence thresholds. It helps evaluate the trade-off between precision and recall. An ideal PR Curve remains high across most thresholds, showing that the model achieves both high recall and high precision

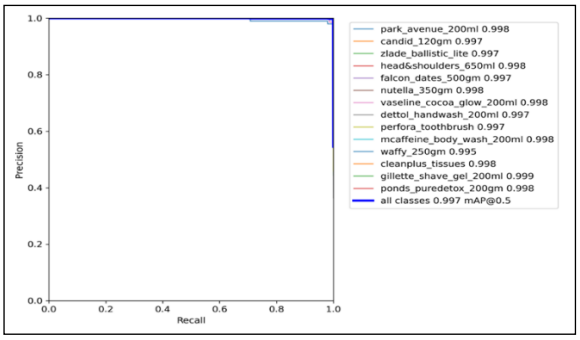

Fig. 6 PR Curve

5) *Model Performance:* It plots a very detailed summary of the YOLOv7 model's performance at each training epoch. Each subplot represents a different characteristic of the learning, where a certain set of metrics evolve with the time (or epochs) when trained on this data and validated on other data. The gradual reduction in losses (Box, Objectness, and Classification) indicates that the model is effectively learning to localize and classify objects. Clearly, the high mAP values indicate that the model is doing a good job in detecting objects with both accuracy and precisions.

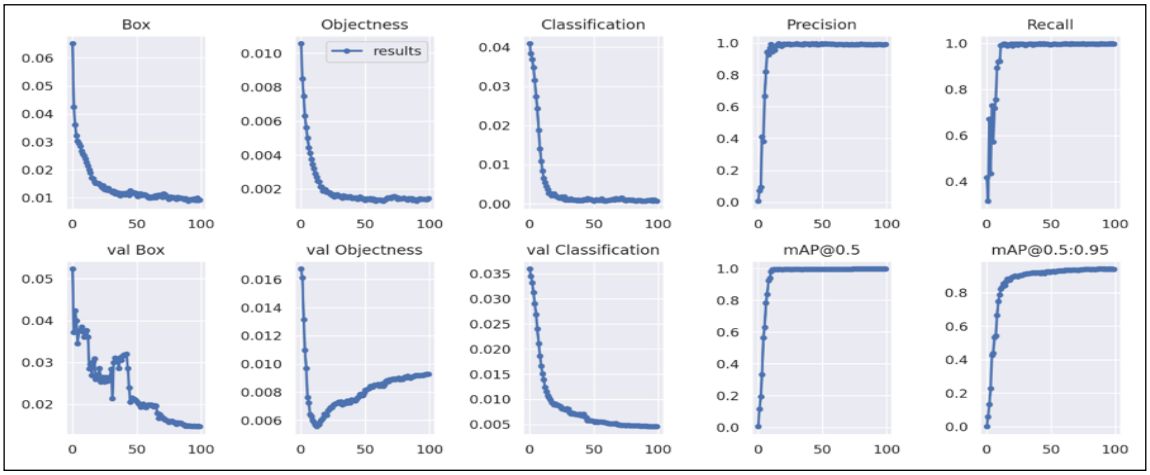

Fig. 7 Model Performance

B. Analysis:

Label category Evaluation Definition	mAP	recall	TP	FP	FN	Average Confidence of detection	Detection Count	Ground Truth Count
cleanplus_tissues	1	1	140	0	0	0.964188	140	140
candid_120gm	1	0.99	214	0	3	0.954284	214	217
dettol_handwash_200ml	1	0.98	192	0	3	0.951258	192	195
falcon_dates_500gm	1	0.98	129	0	3	0.973444	129	132
gillette_shave_gel_200ml	1	0.96	193	0	7	0.960571	193	200
head&shoulders_650ml	1	0.94	209	0	14	0.931982	209	223
mcaffeine_body_wash_200mlm caffeine_body_wash_200ml	1	0.97	169	0	5	0.950048	169	174
nutella_350gm	1	0.98	240	0	5	0.968431	240	245
park_avenue_200ml	1	0.95	162	0	8	0.940611	162	170
perfora_toothbrush	1	0.99	262	0	2	0.974572	262	264
ponds_puredetox_200gm	1	0.98	175	0	3	0.971967	175	178

vaseline_cocoa_glow_200ml	1	0.99	198	0	3	0.959993	198	201
waffy_250gm	1	0.94	215	0	13	0.949646	215	228
zlade_ballistic_lite	1	0.95	106	0	5	0.963752	106	111
Overall	1	0.97	2604	0	74	0.958196214	2604	2678

Table. 1 Model Evaluation

The evaluation metrics for the model YOLOv7 were calculated with an IoU threshold of 0.45 and a confidence threshold of 0.5, which fairly balances detection sensitivity and precision. The model hit the mean Average Precision (mAP) of 1.0 for each of the product categories at these configurations, which meant it would detect and classify objects with considerable overlap with ground truth boxes even at a moderately low IoU threshold very accurately. Thus, the approximate recall score of 0.97 shows that it was able to pick most of them though a few got left out as evidenced in the rather lower recall values for certain product categories such as "head&shoulders_650ml" at 0.94 and "waffy_250gm" at 0.94. Metrics show the model can certainly count on spotting products more than real conditions are guaranteed to leave at perfect overlapping with bounding boxes, thus taking slightly broader detections that improve sensitivity over specificity.

For the confidence threshold at 0.5, the model's predictions are shy but accurate. With the average level of model confidence to be about 0.96 on all the detections, it reduces FP cases wherein no false positive case is detected by the model with zero FP. The number of total detection counts presents very minor differences from the counts of ground truths that further solidified the success of the model in object detection.

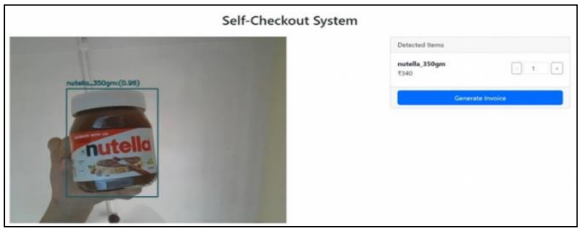

Fig. 8 User Interface

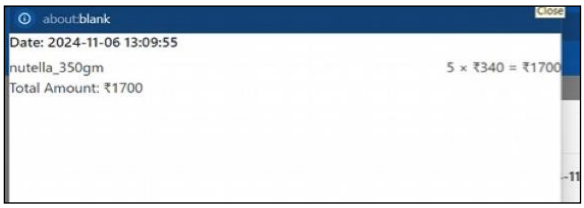

Fig. 9 Invoice Depiction

As shown in the figure 8, this is the visualization of our API application model which is developed using flask in the python framework. In this, we have tested this API model by testing one of our products, "Nutella" which is showing accuracy of 0.96 and automatically gets added to the invoice on the right side. We can place minus or subtraction signs on either side of the place where quantity is stated to alter the quantity. Once we complete the detection process we can print out the invoice generated which can be clearly shown in figure 12, the amount to be paid and quantity of the product.

V. APPLICATIONS, ADVANTAGES & LIMITATIONS

C. Applications

This system is highly suited for retail environments, where it enhances operational efficiency and improves the customer experience by automating checkout processes. It facilitates contactless billing and inventory management, making it an ideal solution for supermarkets and department stores. Beyond retail, the system's object detection and real-time tracking capabilities can be applied across various industries. In warehouses, it can streamline inventory management by ensuring accurate stock counts and timely replenishment. Similarly, in libraries, it can automate the checkout process for books, while in logistics, it can enable efficient parcel scanning and tracking. These examples underscore the system's adaptability and its potential to revolutionize operations in sectors where precision and efficiency in item recognition are critical.

D. Advantages

The CounterVision system powered by real time object detection and tracking, based on the YOLOv7 model has transformational potential for retail environments. The system excels in simultaneously identifying multiple products without reliance on barcodes, which reduces checkout times and improves overall store efficiency. High-speed recognition goes hand in hand with robust tracking capability, ensuring effective operation even in cases of occlusions or dynamic movements about items. In the process, it automates the checkout process; offers several avenues to reducing human error; streamlines inventory management; and ensures maximum deployment of employees to different tasks that require interpersonal contact with customers. It also

meets the expectations of modern consumers in their contactless experience, reductions in waiting time and check complexity, hence increasing customer satisfaction.

A. Limitations

Despite the advances, the CounterVision system has challenges that limit its effectiveness at times. The major limiting factors are exposure to environmental aspects such as varying lighting conditions and product occlusions, all of which negatively impacts detection accuracy. This system's reliance on high-performance computing infrastructure in the form of high-performance GPUs may limit it to usage by only retailers with such infrastructure. Besides, the performance of the system would largely depend upon the quality and diversity of the training data. Therefore, an elaborate dataset is quite imperative for reliability over diverse retail scenarios. Scalability would also be an issue in high volume retail environments since small delays in the real-time detection amplify under heavy loads. Last but not least, the implementation costs, comprising hardware acquisition and training, are against the wide-scale adoption particularly by the small and medium-sized enterprises.

VI. FUTURE SCOPE

The "CounterVision" system presents many opportunities for future improvement, focusing on object detection capabilities and retail-specific applications. Eventually, it will be expanded on by extending the training model with an even more significant and diverse dataset, thus making it capable of better detecting products at different angles and orientations, including non-standard ones, or complex retail environments. There are many potential domain-specific optimizations of YOLOv7, from handcrafted anchor box designs to extra detection heads that take into account gains in precision when detecting oddly shaped or packaged merchandise. Such preprocessing innovations, such as improved lighting normalisation or dynamic region cropping, could alleviate environmental variability and result in more robust and stable performance in deployment scenarios.

Deployment onto hardware like NVIDIA Jetson would shift the computational intensive load out and make the system viable even for smaller, less infrastructurally complex retail environments. Adaptive learning mechanisms leveraging operational data in real-time may improve detection accuracy overtime with incremental updates to models. Even other deployments of the system in retail could involve integration with other systems, such as inventory management and analytics, that can create greater value for retailers. These technological advances would place "CounterVision" in such a way as being scalable, efficient, and adaptive to new challenges that the evolving retail industry is facing.

VII. CONCLUSION

The "CounterVision" system is the epitome of retail automation through the use of state-of-the-art computer vision techniques for real-time object detection. Its model, YOLOv7, is optimized for the balance between speed and accuracy, which condition needs to be met for the actualization of its robust multi-object detection capability in complex retail scenarios. For technical fine-tuning, it adopts the use of hyperparameters such as learning rate, 0.001; momentum, 0.937; and weight decay, 0.0005 for outstanding performance to be achieved. The system achieved a mean Average Precision of 1.0 and recall of 0.97 at a confidence threshold of 0.5. It makes use of only one fixed-camera setup and data preprocessing techniques, in the form of background subtraction and mosaic augmentation, to reduce computations yet maximally maintain high detection accuracy. The annotated dataset that is built using the tool called LabelMe ensures proper handling of variable lighting as well as diverse product appearances. "CounterVision" does not rely on barcodes, nor does it require manual scanning, thereby smoothing the checkout process that is contactless, efficient, and user-friendly. The model remains open to challenges such as scaling, real-world adaptability, and environmental variations for further development.

REFERENCES

[1] Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies methods for real-time object detectors. 2024. DOI: https://doi.org/10.1109/CVPR52729.2023.00721

[2] Avanzi, A., Brémond, F., Tornieri, C. et al. Design and Assessment of an Intelligent Activity Monitoring Platform. EURASIP J. Adv. Signal Process. **2005**, 318538 (2005). DOI: https://doi.org/10.1155/ASP.2005.2359

[3] Pal, Sankar & Pramanik, Anima & Maiti, Jhareswar & Mitra, Pabitra. (2021). Deep learning in multi-object detection and tracking: state of the art. Applied Intelligence. 51. DOI: https://doi.org/10.1007/s10489-021-02293-7

[4] Xia, K.; Fan, H.; Huang, J.; Wang, H.; Ren, J.; Jian, Q.; Wei, D. An Intelligent Self-Service Vending System for Smart Retail. Sensors 2021, 21, 3560. DOI: https://doi.org/10.3390/ s21103560

[5] Xia, Kun, Hongliang Fan, Jianguang Huang, Hanyu Wang, Junxue Ren, Qin Jian, and Dafang Wei. 2021. "An Intelligent Self-Service Vending System for Smart Retail" Sensors 21, no. 10: 3560. DOI: https://doi.org/10.3390/s21103560

[6] B. -F. Wu, W. -J. Tseng, Y. -S. Chen, S. -J. Yao and P. -J. Chang, "An intelligent self-checkout system for smart retail," 2016 International Conference on System Science and Engineering (ICSSE), Puli, Taiwan, 2016, pp. 1-4, DOI: 10.1109/ICSSE.2016.7551621.

[7] Lukezic A, Vojir T, ˇCehovin Zajc L, Matas J, Kristan M. Discriminative correlation filter with channel and spatial reliability. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 6309-6318). DOI: https://doi.org/10.1007/s11263-017-1061-3

[8] Ellis, Griffin, "Automated Retail Checkout by Computer Vision: ARC-AI" (2023). Computer Science and Engineering Senior Theses. 248.

[9] Bewley, Alex & Ge, Zongyuan & Ott, Lionel & Ramos, Fabio & Upcroft, Ben. (2016). Simple online and realtime tracking. 3464-3468. DOI: https://doi.org/10.1109/ICIP.2016.7533003.

[10] Koeferl F, Link J, Eskofier B. Application of SORT on Multi-Object Tracking and Segmentation. InIEEE Conference on Computer Vision and Pattern Recognition Workshops 2020.

[11] Guimarães, V., Nascimento, J.; Viana, P.; Carvalho, P. A Review of Recent Advances and Challenges in Grocery Label Detection and Recognition. Appl. Sci. 2023, 13, 2871. DOI: https://doi.org/10.3390/ app13052871