Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4916818, 18 pages
https://doi.org/10.1155/2022/4916818

Research Article

Hybrid Approach for Shelf Monitoring and Planogram Compliance (Hyb-SMPC) in Retails Using Deep Learning and Computer Vision

Mehwish Saqlain,[1] **Saddaf Rubab**,[1,2] **Malik M. Khan**,[1] **Nouman Ali**,[3] **and Shahzeb Ali**[4]

[1]*National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan*
[2]*Department of Computer Engineering, College of Computing and Informatics, University of Sharjah, Sharjah 27272, UAE*
[3]*Department of Software Engineering, Mirpur University of Science and Technology (MUST), Mirpur 10250, (AJK), Pakistan*
[4]*COMSATS University, Islamabad 44000, Pakistan*

Correspondence should be addressed to Saddaf Rubab; saddaf@mcs.edu.pk

Received 4 February 2022; Accepted 10 May 2022; Published 15 June 2022

Academic Editor: Abdul Qadeer Khan

In retail management, the continuous monitoring of shelves to keep track of the availability of the products and following proper layout are the two important factors that boost the sales and improve customer's level of satisfaction. The studies conducted earlier were either performing shelf monitoring or verifying planogram compliance. As both the activities are important, to tackle this problem, we presented a deep learning and computer vision-based hybrid approach called Hyb-SMPC that deals with both activities. The Hyb-SMPC approach consists of two modules: The first module detects fine-grained retail products using one-stage deep learning detector. For the detection part, the comparison of three deep learning-based detectors, You Only Look Once (YOLO V4), YOLO V5, and You Only Learn One Representation (YOLOR), is provided and the one giving the best result will be selected. The selected detector will perform detection of different categories of SKUs and racks. The second module performs planogram compliance; for this purpose, the company-provided layout is first converted to JavaScript Object Notation (JSON) and then the matching is performed with the postprocessed retail images. The compliance reports will be generated at the end for indicating the level of compliance. The approach is tested in both quantitative and qualitative manners. The quantitative analysis demonstrates that the proposed approach achieved an accuracy up to 99% on the provided dataset of retail, whereas the qualitative evaluation indicates increase in sales and customers' satisfaction level.

1. Introduction

Retailing encompasses the selling of goods and services. It is an integral part of the modern society and also acts as a driving force by contributing significantly to the GDP and aims at encouraging sustained growth [1]. The improvement of living standards of society leads to evolution of retails at an accelerated rate. As Artificial Intelligence (AI) is revolutionizing every sphere of life, enterprises are also focusing on using AI to reshape the ecology of retail industry [2]. Retail management and improving customer's experience are a challenging task due to multitude of tasks required to be performed in concurrent manner, for example, inventory management, shelves organization, and customer's support.

The predefined arrangement of products within shelves is called *planogram*: it demonstrates the layout of placing each product within shelves and indicates how many *facings* should be present, that is, how many stock-keeping units (SKUs) of the same product should be visible in the front row of the shelf [3]. The effective organization of SKUs on the shelves attracts more customers and helps them to choose and pick the products in an efficient manner [4]. To achieve this objective,

corporations invest in tools and studies to create planograms that are part of their optimal store policy. After proper planning, a layout of placing stock-keeping units (SKUs) is decided by the headquarter, and that particular layout is communicated to the retailers. Retailers are then offered certain discounts or monetary benefits for following the planogram provided to them. As the organizations are investing time and money, they also need to verify the compliance of their planogram by the retailers and stores. At present, the verification of planogram compliance is the responsibility of the store personnel and is routinely performed [5].

The auditing of the shelves, that is, keeping track of SKUs availability, their number, and positioning at different locations, is necessary for optimized management of the retails. The expansion of retail industry with the passage of time also makes shelf monitoring a tough job. Traditionally, the auditing of the shelves is performed manually by store representatives in retail environment. The manual approach is very time-consuming, requires extensive human labor, and is subject to human errors. All these aspects contribute to making inventory management a difficult task.

The important factors that boost sales and improve satisfaction level of customers are (a) the availability of products and (b) the arrangement of products on supermarket shelves [6]. One of the major problems in retail environment is out-of-stock products; a study conducted in [7] showed that, in case of no availability of the required products, 31% of the customers prefer to move to other stores, 26% switch to another brand, and 15% delay their purchase to some other time, whereas 9% buy nothing. This illustrates that on-time availability of the products is a crucial factor affecting the sales environment.

Research also indicates that following 100% optimal planogram can amplify the sales to 7.8% and boost the profit [4]. It also helps merchandisers to make more effective decisions about inventory management. The management of proper counters of available products and producing alerts in case of misplaced SKUs and decreasing the levels of products will encourage the organizations to take appropriate steps and decrease the stocking issues.

For the optimized retail management, planogram compliance and shelf monitoring should be performed in an automated way. To automate this process, object detection in the images of shelves can solve problem of monitoring different categories and subcategories of SKUs, completing missing SKUs, and matching planogram continuously. Fine-grained object recognition in retail industry is a challenging task due to below-mentioned reasons.

Racks are not properly organized and variation in product poses causes problems; products are placed in different order; they are often placed in a horizontal manner. This cluttered condition causes complexity of scene as depicted in Figure 1(a). Different resolution of image capturing device produces different quality images, making product detection difficult. In different strategies of capturing images and variation in the image parameters, the length of different products is mapped to different resolution of pixel (Figure 1(b)). The jitter and camera shake while capturing images cause blurry images which make it difficult to recognize the products as the details are not clearly visible to be detected (Figure 1(c)).

The catchy, glary, and glossy packages of products and uneven illumination, shadows, and lightening conditions cause reflection as illustrated in Figures 1(d) and 1(e). The images captured from the oblique angles and not from the frontal view cause distortion (Figure 1(f)). Different shapes, sizes, colors, and minute visual difference in product packages require fine-grained classification (Figures 1(g) and 1(h)).

Due to all these problems, fine-grained product recognition becomes difficult. The analysis of existing studies indicated that the studies conducted in the past were either performing shelf monitoring or checking planogram compliance. As both activities are critical in optimized management of retails, a hybrid approach is required to perform both activities in an efficient manner. To deal with this challenge, a hybrid approach for shelf monitoring and planogram compliance which is called Hyb-SMPC is proposed in this work. For the detection part, three deep learning models (YOLO V4, YOLO V5, and YOLOR) will be compared and the one having accurate results will be used. The following are the contributions of this research work:

(i) Fine-grained SKUs detection in the retail by using deep learning.

(ii) Shelf monitoring, that is, keeping track of the instances of the SKUs present on the shelves.

(iii) Verifying the compliance of the planogram using techniques of computer vision.

The organization of the rest of this paper is as follows. Section 2 presents related work. Section 3 provides the details of the proposed approach. Results are provided in Section 4 and, finally, the conclusion and the future work are provided in Section 5.

2. Related Work

2.1. State of the Art in Computer Vision. In our daily life retails are playing a significant role. The number of products in the retail increases every day with the increasing number of new products coming into the market; in this situation, the traditional way of retail management is very difficult as product detection and inventory management require extensive human labor.

Deep learning and computer vision are used in various applications such as image classification, object detection in industrial production [8], medical image analysis, and action recognition [9, 10]. To automate this manual process, various computer vision-based solutions have been proposed in the literature. The first attempt of product recognition in grocery stores was done in [11], where the authors applied three different object recognition algorithms based on local invariant features. The proposed methodology did not perform well considering precision and efficiency but one of the important contributions of that work was the provision of a dataset called GroZi-120 consisting of 120 different grocery products. The dataset is publicly available for further research [11]. Feature-based product detection methods

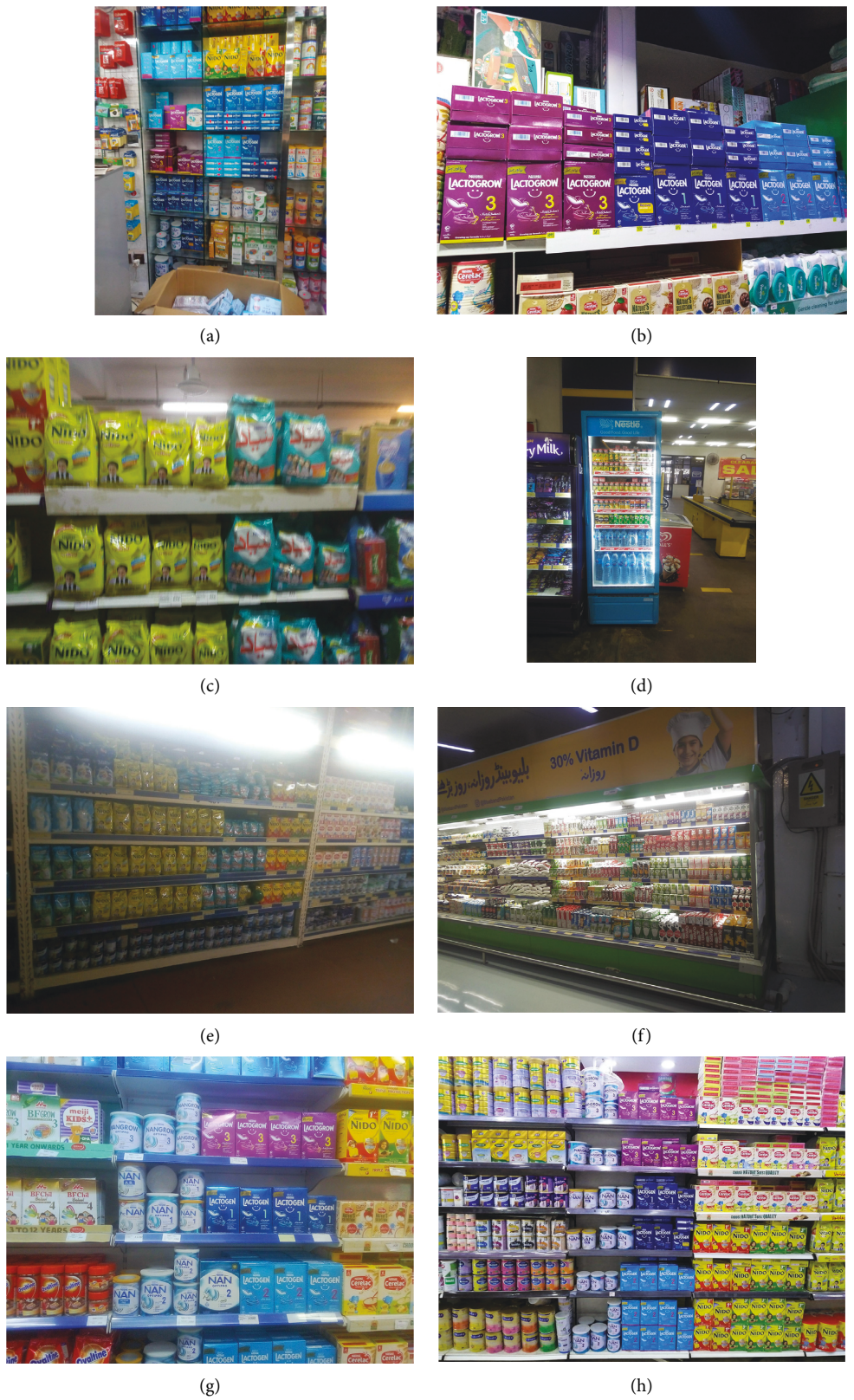


Figure 1: Improper product placements on the racks. (a) Unorganized products. (b) Difference in dimension. (c) Blurry image. (d) Uneven illumination. (e) Lightening reflection. (f). Nonfrontal image. (g). Minute difference of products. (h) Change in product size.

were extensively studied in the past. These features are key-point-based, gradient-based, color-based, pattern-based, and deep-learning-based [12].

2.1.1. Key-Point-Based Features. Key-point-based feature detection method detects the key points from the images, and this is the technique frequently employed in the retail scenario. Scale-Invariant Feature Transform (SIFT) [13] and Speeded-Up Robust Feature (SURF) [14] are the well-known feature extraction algorithms. SIFT features are invariant to scale, rotation, and illumination of image. By using SIFT in [15], features of input images were compared with stored features of the objects in the database and recognition was performed; however, in [16], the on-shelf availability and misplaced products were detected by using SURF descriptors. In the first phase, counting is performed; by looking at the duplicate properties provided by SURF descriptor, each item of the product is counted. In the second phase, rectangular bounding boxes are fitted around each product and the products are identified with SURF and color properties.

Reference [17] employed logo-detection-based algorithm for the recognition of products on shelves. The algorithm includes two steps. Products were detected and classified on the basis of their brands by matching SIFT key points and the finer classification was performed by using color information and the exact product label was recognized. Visual monitoring of shelves was performed in [4]; the system provided in the study analyzed the images and verified compliance of the planogram. For the detection of the objects, three different approaches were used; the vote map approach used SURF descriptors and outperformed the other approaches.

2.1.2. Gradient-Based Features. These features focus on the structure and shape of the object. Histogram of Oriented Gradients (HOG) and Sobel and Prewitt operators are generally used for gradient computation. These operators incorporate geometric shapes, edges, and corners of the products detected from the images [12]. Gradient-based features like color and shape were captured in [18] for the purpose of product detection in retails. Sobel operator was used in [19] and the authors presented an automated planogram compliance check in retails by providing a framework based on visual analysis. The framework consists of three modules which performed row extraction and occupancy computation and identified completely and partially missing case. Through exploiting color and texture properties, the counting of the products was performed, and their placement was checked. Nevertheless, the study did not perform product detection. In [20], a heuristic approach to count the instances of the same product and detect the missing item on the shelf without using classifiers is specified. The proposed algorithm includes morphological operations, template matching, and histogram comparison. The experimental results demonstrate that the satisfactory results are achieved with the algorithm only when the user manually selects the most substantial label of the product from the shelf image.

2.1.3. Pattern-Based Features. In the detection of retail products using recurring patterns, Haar and Haar-like features are extensively used pattern-based features [12]. A fine-grained recognition of grocery products by integrating VGG-16 with recurring features and attention maps was proposed in [21]. Recurring features detect the candidate region and give coarse labels to the products; afterwards attention maps help the classifier to concentrate on the fine details in the candidate region of interest (ROI). The mean average precision (mAP) of 0.75 is achieved with the proposed method. The authors in [11] analyzed the performance of boosted Haar-like features, SIFT, and color histogram matching algorithms and found that SIFT outperformed the others. An automatic method for checking the planogram compliance is suggested in [22] without the requirement of template images through the detection of recurring features. Graph matching technique was used to match the provided layout with the extracted layout.

All of the above-mentioned studies conducted for object detection and planogram compliance used traditional computer vision techniques. These traditional techniques require manually designed features; these hand-crafted features do not always reflect sufficient information. Each feature characterization requires working with a plethora of parameters [23]; with increasing number of classes, feature extraction becomes inconvenient and it becomes the responsibility of CV engineers to select the features which identify different classes of objects in the best manner. To deal with this difficulty, deep-learning-based methods were introduced, which are based on the concept of end-to-end learning [23]. Recent research is focused on the use of mid-level features and deep learning models to build robust decision support systems such as smart vehicles and IoT applications [24–28].

2.2. State of the Art in Deep Learning. Over the time, deep learning has emerged efficiently by showing improved performance. Deep learning has the ability to learn the features automatically from the images [23]. Another merit of deep learning is the deeper layers, which can extract precise features, whereas simple neural networks are not able to do that [2]. The most frequently used technique in deep learning is convolutional neural network- (CNN-) based object detection. You Only Look Once (YOLO) [29], Single-Shot Detector (SSD) [30], and Region-based Convolutional Neural Networks (R-CNN) [31] are the modern variants of CNN which are very efficient. CNN outperformed traditional methods based on hand-crafted features such as SIFT [13] and SURF [14] as they are unable to extract deep information from images.

Many researchers contributed to using CNN for detection of products in retails. Reference [32] used convolution neural network to resolve the issue of in-store product recognition and achieved an accuracy of 78.9%. The challenge of large-scale fine-grained structure classification was handled in [33] by exploiting contextual information along with deep network.

To investigate the number of products present on the shelf, [34] took the help of surveillance cameras to record the

videos of the shelves to take account of the number of the present products. The study tracks the changed regions by background subtraction method; afterwards moving objects are removed and CNN based on CaffeNet is used for the classification of changed regions. The success rate of 89.6% was achieved with this study [34]. The extension of the work is provided in [35] which used images from the surveillance camera for monitoring the availability of products. The Hungarian method distinguishes the foreground from the successive image. The classification of detected changed region is performed by two deep networks, that is, CIFAR-10 and CaffeNet. This methodology also helps in the determination of shelves which are accessed commonly [35]. A fast detection and recognition method based on fine-grained categories of products is anticipated in [36] when very limited training data is available for training. The results indicate that 52.16% mAP was achieved for recognition of each product.

Reference [37] provided a template-free, zero-short product detection system which avoided templates and detected the products by segmenting the shelves horizontally into layers and vertically into products. The classifications of horizontal layers are performed by GoogLeNet, whereas vertical division is performed by another trained GoogLeNet. The results indicate better performance compared to the existing methods; however, the empty regions between the products influence the method negatively, making it less robust.

A deep learning approach was suggested by [38] for planogram compliance in retail stores. The images are collected through the robot NAVii and also from the Internet and then split into three different training sets for training three different CNN models. The CNN model trained on both Internet and store images gives better accuracy than other models and can generalize in a much better way because of exposure to the great variation of products.

Deep-learning-based object detection methods are divided into two categories: two-stage detectors and one-stage detectors. Region-based Convolutional Neural Networks (R-CNN) [39], Faster R-CNN [31], and Mask R-CNN [40] come under the category of two-stage detectors. One-stage detectors deal with object detection as a simple problem of regression. RetinaNet [41], YOLO [29], and SSD [42] are well known one-stage detectors.

Reference [43] provided a semisupervised deep-learning-based image classification approach for shelf auditing. The study merged the two ideas of "semisupervised" and "on-shelf availability (SOSA)." Semisupervised learning took advantage of both labeled and unlabeled data. Deep learning architecture YOLO V4 is used for on-shelf auditing (OSA); it makes comparison of three different approaches of deep learning (RetinaNet, YOLO V3, and YOLO V4) for monitoring OSA and the best results were achieved with YOLO V4; however, the study did not perform planogram compliance.

There are very few studies regarding checking of planogram compliance in retails. In [44], deep-learning-based hybrid approach based on image classification and object detection is provided to solve the problem of planogram compliance in retails. For assessment of quality of the images, Blind/Referenceless Image Spatial QUality Evaluator BRISQUE [45] technique was integrated into the framework. Eight different types of templates were taken into account to train the model. The products were classified into two classes, that is, "Exact 7 by 4" and "No Exact 7 by 4." VGG-16 was used for classification and Tiny YOLO V2 [46] was used for object detection. The overall accuracy achieved by this hybrid approach reaches 95% [44].

YOLO V5 and YOLOR are recently released versions, so no work has yet been done in the retail industry using these models. The proposed study is the first one to provide comparison of these models in shelf monitoring and planogram compliance. The summary of the studies conducted in the past is given in Table 1 which gives the clear idea that all the studies conducted in the past were performing one of two tasks: shelf monitoring and planogram compliance. Hence, it was discovered that there is a dire need for a hybrid approach that can perform both shelf monitoring and planogram compliance in retails. The novelty of the proposed approach is to perform both tasks.

3. Proposed Approach: Hybrid Approach for Shelf Monitoring and Planogram Compliance (Hyb-SMPC)

The proposed technique is a hybrid approach which combines both concepts, "shelf monitoring" and "planogram compliance," for the first time to facilitate retail management. The process involves object detection at fine-grained level followed by verification of planogram compliance using shelf images. For this purpose, the study used three one-stage deep learning detectors, that is, YOLO V4 [47], YOLO V5 [48], and YOLOR [49], to detect and classify the products. The proposed study is the first one to provide comparison of these three detectors. The approach is broadly divided into two modules as illustrated in Figure 2.

The first module is product detection module. It performs detection and localization of racks as well as SKUs in parallel manner. The process is followed by the next module called planogram compliance module, which verifies the specific placement policy of SKUs formulated by the company.

The general process of training the models is represented in Figure 3. The study used image-based dataset which is provided by the industry partner and is collected from different retails, stores, and supermarkets. As the dataset is image-based, preprocessing involves images resizing, denoising, and image labeling. Image labeling is a process of providing annotations to the images. Labels are provided on the basis of type of products. Different retails contain variety of racks, so racks are also labeled in the preprocessing stage. After preprocessing stage, the next phase is the splitting of data into two subsets, that is, training and testing, and then training of three different detectors is performed using labeled data of training subset. For training 200 images per product, SKUs are used.

Table 1: Comparison of proposed approach with the previous approaches.

Reference no.	Year	Object detection			Planogram compliance	Methods
		Traditional method	Deep learning			
			Two-stage	One-stage		
[4]	2015				✓	SURF
[16]	2015	✓				SURF
[18]	2015		✓			SVM
[19]	2015				✓	SURF + color histogram
[22]	2016				✓	Recurring pattern detection
[38]	2016				✓	CNN
[36]	2017		✓			VGG-F
[17]	2018	✓				SIFT
[21]	2018		✓			VGG-16 with recurring features and attention maps
[37]	2018		✓			GoogLeNet
[35]	2019		✓			CIFAR-10, CaffeNet
[43]	2020			✓		RetinaNet, YOLO V3, YOLO V4
[44]	2020				✓	VGG-16, Tiny YOLO V2
Proposed method				✓	✓	YOLO V4, YOLO V5, YOLOR

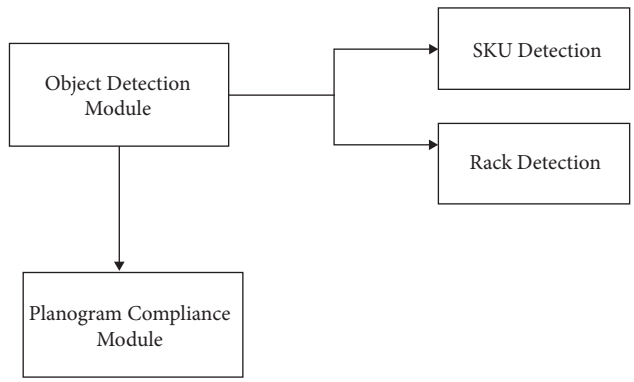


Figure 2: Modular view of Hyb-SMPC.

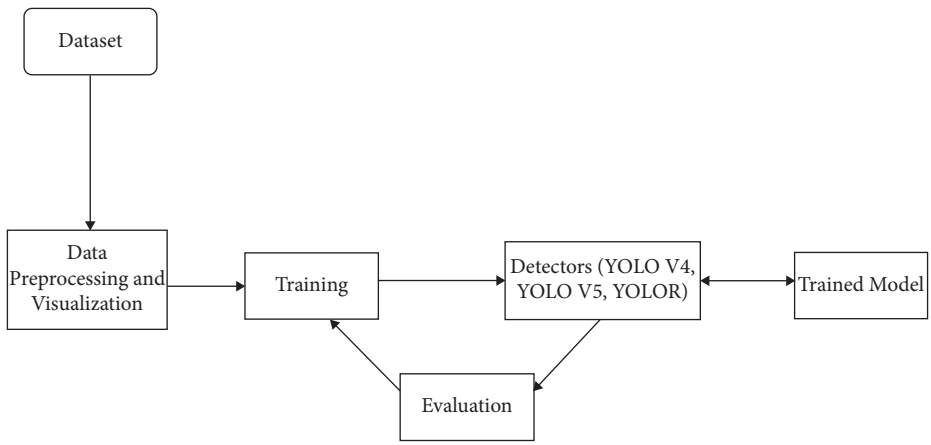


Figure 3: Model training process.

Three one-stage detectors are trained by providing labeled data (YOLO V4, YOLO V5, and YOLOR). The SKUs of the dataset are categorized under twelve categories which are mentioned in Table 2. The testing of the detectors is performed with the test set and performance is evaluated using different accuracy metrics.

After training and tuning of detectors are performed, the detector producing the most accurate results among the three will then be deployed on the server to perform detection. The detailed flow of the proposed hybrid approach is represented in Figure 4. The first step is the provision of labeled images from the real retail environment to the three detectors for training. The next step is the selection of the best model depending upon the results of different evaluation metrics. Afterwards unlabeled images would be provided to the best trained detector; the selected detector will then perform detections.

After detection, the detector provides the processed images as indicated in Figures 5(a) and 5(b) which contain the list of detected SKUs, along with their counters and IDs to illustrate the number of instances of the SKUs present on the shelves. Figures 5(c) and 5(d) show the racks detected by the detector. Later SKUs and racks are sorted with respect to *x, y* coordinates and postprocessed retail images are generated. The sorting of SKUs and racks with respect to coordinates is very important as it will act as an input for the second module and provide help in planogram checking. The postprocessed retail images illustrated in Figures 6(a) and 6(b) contain sorted SKUs in the sorted racks.

After the detection of SKUs and racks, the second module called the planogram compliance module will start its working. For planogram matching, the study explored two methods:

(1) Planogram matching using color detection

(2) Planogram matching by generating JSON

3.1. Planogram Matching Using Color Detection. In planogram matching using color detection for verification of layout, the postprocessed retail images received from the first module are converted into planogram images by using Python libraries called OpenCV and PIL image. Afterwards, the generated planogram and the company-provided planogram are matched.

For making planogram image, all the racks, shelves, and SKUs will be represented as 2D blocks. A rectangular block which depicts the whole shelf is first generated (yellow block in Figure 7(b)) by taking into account the shape of a retail shelf. The racks are then generated one by one (green box in Figure 7(b)) by using the information contained in the postprocessed retail images. Each rack contains different two-dimensional (2D) boxes filled with colors representing different categories of SKUs. Different colors are assigned to different categories of SKUs as depicted in Figures 7(a) and 7(b). The postprocessed retail image obtained from the first module is represented in Figure 7(a), whereas Figure 7(b) represents the planogram image generated from it. The company-generated layout is given in Figure 7(c) as a template. We are using a color dataset that contains Red, Green, Blue (RGB) values with their corresponding names. The CSV file for the color dataset has been taken from [50]. The dataset file includes 865 color names along with their RGB and hex values. Now we assign each RGB color value to the classes belonging to each category. Hence, each color represents a specific SKU in the planogram.

Matching is performed on the basis of same image size and the same number of racks as the company-provided template; for example, the planogram of category chiller with four racks will only be matched with the planogram image of chiller with four racks generated by our planogram module. As we have all the coordinates of racks and SKUs provided by detectors, we pick the cropped image part of each SKU from both planogram images (company-provided and module-generated) to verify its color. If the RGB values of both cropped images get matched, then we store TRUE as a string in an object. Similarly, if they cannot be matched, then we store FALSE.

Afterwards, by counting the number of total TRUEs and FALSEs, we will generate the report of planogram compliance. The threshold values are decided as 10%–90%; a value lying between these limits indicates that the planogram is followed partially, whereas a value above 90% indicates planogram to be followed fully and a value below 10% indicates that planogram is not followed at all.

3.2. Planogram Matching by Generating JSON. In this method, the company-provided template which is shown in Figure 8(b) is given as an input and the module will generate JSON from it using Python functions. These functions will extract racks and SKUs. The generated JSON will be matched with the information extracted from postprocessed retail image (Figure 8(a)) which is also saved in the form of JSON to find whether the planogram is followed fully or partially or is not followed at all.

The matching of company-generated layout with postprocessed retail images of shelves occurs at real time so this process must be efficient. Hence, for this purpose, we used JSON for matching. The matching occurs rack by rack and SKU by SKU starting from the rack one. When the string of both JSONs gets matched, we store TRUE, and if they cannot be matched, we store FALSE as a string in another object. Threshold values are decided as 10%–90%; a value lying between these limits indicates that the planogram is followed partially, whereas a value above 90% indicates planogram to be followed fully and a value below 10% indicates that planogram is not followed at all. The report is generated at the end and is sent to the company as well as to the retailers. Figure 8(a) indicates that the planogram is followed 100%, whereas Figure 8(c) indicates an image which is not following planogram at all.

The comparison of the two methods described above is performed and the average processing time taken for matching planogram for different categories of the products is calculated. Table 3 contains the details. On the basis of the average processing time, it is evident that the planogram matching by generating JSON is more efficient compared to

Table 2: Main categories of dataset.

No.	Main categories
1	Juices
2	Chiller
3	Dairy liquid
4	Dairy powder
5	Coffee
6	Milk modifier
7	PTW (powder tea whitener)
8	Infant nutrition
9	BFC (breakfast cereals)
10	Nutrition
11	Nestrade
12	Sachets

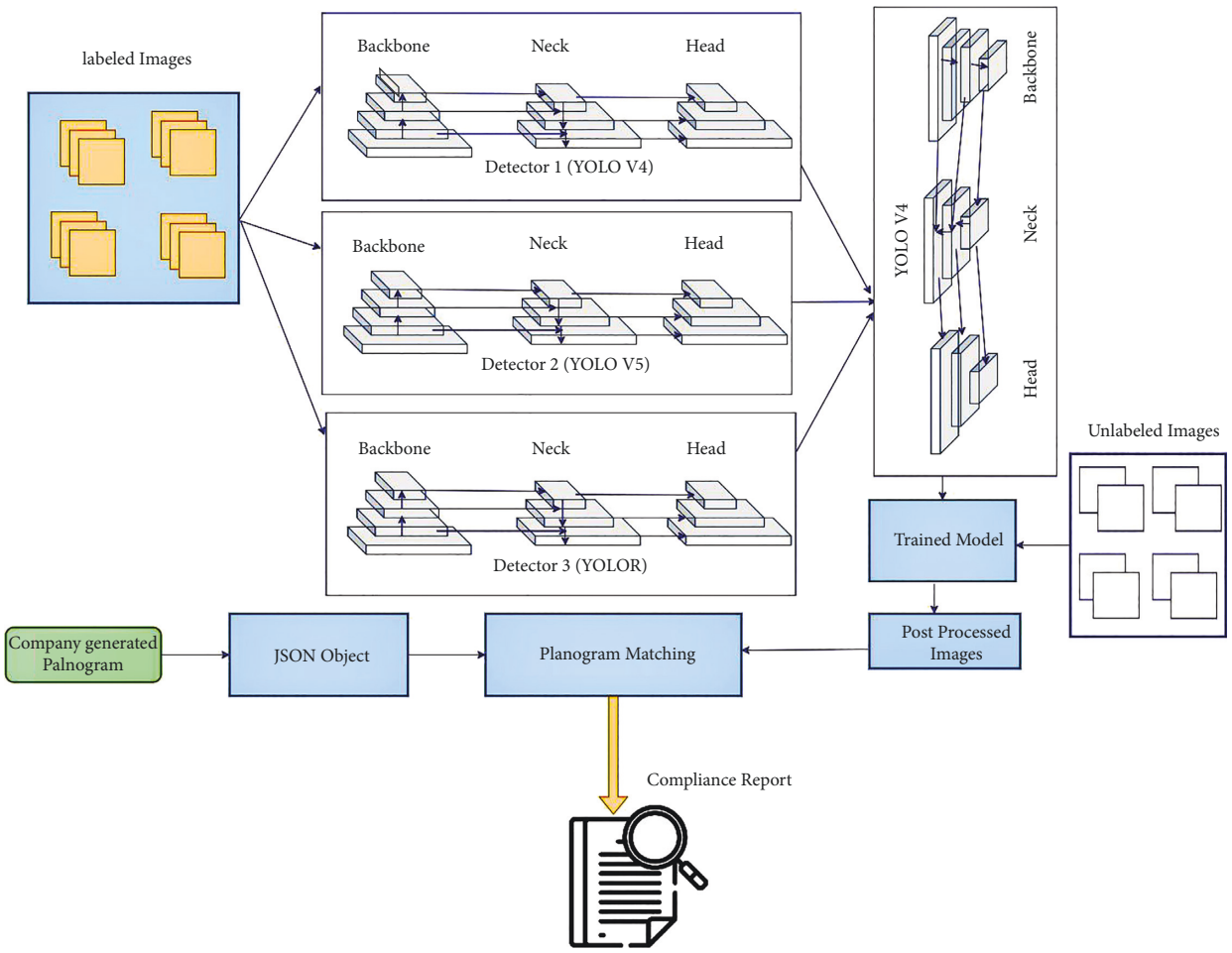


Figure 4: Workflow of Hyb-SMPC.

the other method. Therefore, in Hyb-SMPC, the planogram is matched through generating JSON.

3.3. Formal Definition of Hyb-SMPC Approach. Assume that $D = \{(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots, (a_p, b_p)\}$ has p images with labeled SKUs. In each factor (a, b), $a \in A$, input space, whereas $b \in B = \{l_1, l_2, l_3, \ldots, l_q\}$ has q class labels. The proposed approach considers a function $f: A \leftrightarrow B$ for mapping unseen input images (IM) to correct class labels $(\hat{b})$. $M = \{m_1, m_2, \ldots, m_j\}$, where m refers to one-stage detectors; in the proposed approach, m_1 refers to YOLO V4, m_2 refers to YOLO V5, and m_3 refers to YOLOR; hence, $j = 3$.

The general flow of the proposed approach is given in Algorithm 1 containing five steps. The first step is division of dataset D into D_{train} and D_{test} by 80 percent and 20 percent and training of detectors is performed. In the second step, all the three detectors will be tested using labeled test dataset D_{test} and the best detector is selected by comparing the obtained results. The third step of the algorithm gives the input images (IM) to

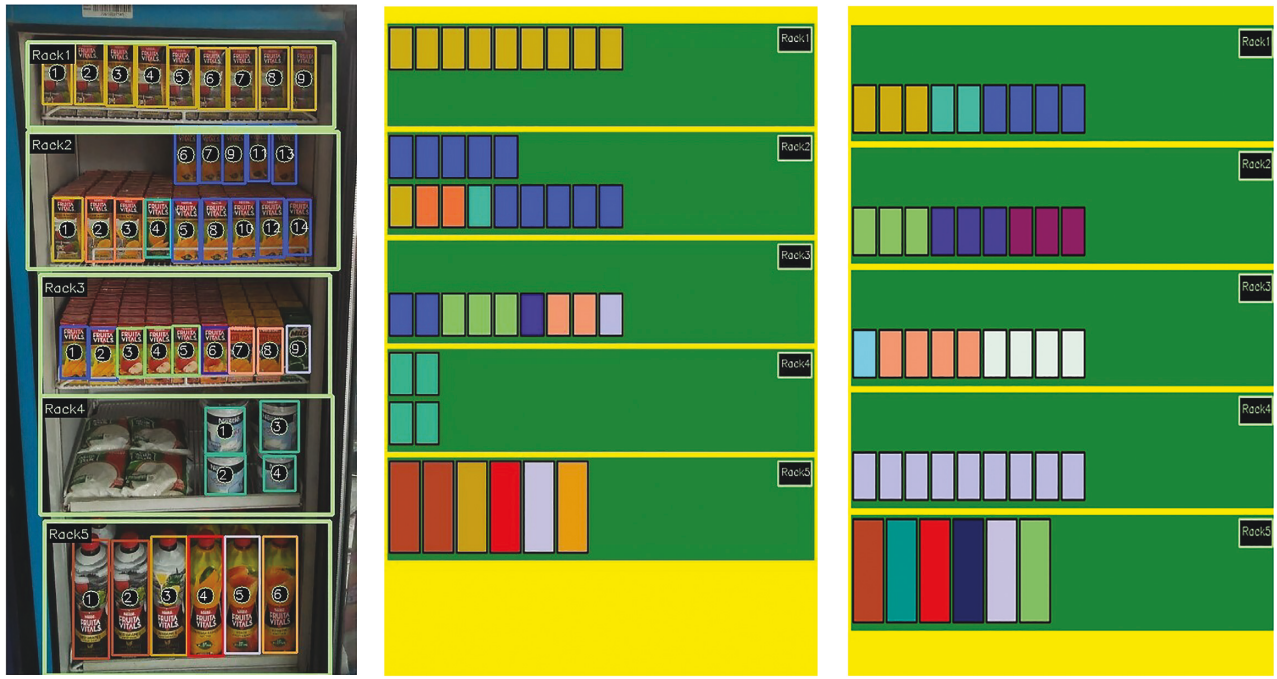
(a) (b)

(c) (d)

Figure 5: Processed images containing ((a) and (b)) detected SKUs and ((c) and (d)) detected racks.

(a) (b)

Figure 6: Postprocessed retail images: (a) postprocessed image 1; (b) postprocessed image 2.



(a) (b) (c)

Figure 7: Planogram matching by color detection. (a) Postprocessed retail image. (b) Generated planogram. (c) Company-provided planogram.

(a) (b) (c)

Figure 8: Planogram matching by generating JSON. (a) Postprocessed retail image. (b) Company-generated planogram. (c) Unmatched postprocessed retail image.

Table 3: Average processing time for planogram matching.

Categories	Planogram matching through color detection (sec)	Planogram matching by generating JSON (sec)
Juices	7	1.5
Chiller	6	1.7
Dairy liquid	7.1	1.2
Dairy powder	7.2	1.3
Coffee	8.2	1.2
Milk modifier	8	1.5
PTW	7.8	1.8
Infant nutrition	6	1.7
BFC	4.9	1.4
Nutrition	5.9	1.5
Nestrade	6.7	1.5
Sachets	7.7	1.9

Input: D: Labeled dataset $D = \{(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots, (a_p, b_p)\}$ with p images
$B = \{l_1, l_2, l_3, \ldots, l_q\}$ with q classes
$M = \{m_1, m_2, \ldots, m_j\}$ with j models
IM = input images
Output: Trained models
$\hat{b}$ = labels of Classes for the SKUs included in input images
Start:
D_{train} = Split (D, $p * 80$)
D_{test} = Split (D, (p-($p * 80$)))
//Step 1—Training of models with labeled data
for $n = 1$ **to** j:
  for every epoch:
    for every (a_i, b_i) **in** D_{train}:
      m_j = Train (a_i, b_i)
    end
  end
end for
//Step 2—Testing models
for $k = 1$ **to** j:
  for every (a_i, b_i) in D_{test}:
    Prediction = m_k (a_i)
  end
  end for
//Step 3—Detecting SKUs in input image$\hat{b}$ = TM (a_i)
Output: Processed images (PI)
//Step 4—Sorting SKUs and Racks
PPI = Sorting (PI)
//Step 5—Generating Planogram from JSON object and comparing post processed image with Planogram layout
JO = contour (Pg)
foreach a_i in DCompare (PPI, JO)
End Algorithm

Algorithm 1: Algorithm for Hyb-SMPC

find the predicted class labels $\hat{b}$ for different SKUs through trained detectors M and produces processed images (PI). In the fourth step, SKUs and racks are sorted with respect to x, y coordinates and postprocessed retail images (PPI) are obtained. In the fifth step, JSON (JO) is generated from company-provided planogram template; this step will also match post-processed retail images (PPI) with JO for checking compliance.

4. Experimentation and Results

Evaluation is the vital part of any system and the performance of the models is generally evaluated through experimentation. Different accuracy metrics were used to gauge the efficiency of the proposed approach. The details are provided below.

4.1. Evaluation Metrics. This study evaluates the approach both quantitatively and qualitatively. For evaluating our approach quantitatively, the metrics of precision, average precision (*AP*), mean average precision (*mAP*), recall, and the value of *F1-score* are used to estimate the accuracy of the models [51].

True Positive (***TP***): Correctly identified the correct SKU.

True Negative (***TN***): Correctly identified that it is not the correct SKU.

False Positive (***FP***): Also called false alarm, identifies the wrong SKU.

False Negative (***FN***): The SKU is not identified when actually it should be identified.

Precision specifies correct detections over total number of detections.

$$\textbf{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}. \quad (1)$$

Recall indicates the number of totally corrected SKUs from the list of SKUs visible in the image:

$$\textbf{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}. \quad (2)$$

*F*1-score merges both precision and recall:

$$\textbf{F1 Score} = 2.\frac{\text{Precision . Recall}}{\text{Precision} + \text{Recall}}. \quad (3)$$

Average precision is calculated by taking average of all the values of precision:

$$\textbf{AP} = \frac{1}{N}\sum_{i=0}^{N} \text{Precision}. \quad (4)$$

Mean average precision is the average of all the APs:

$$\textbf{mAP} = \frac{1}{|T|}\sum_{i \in T}^{N} \text{A}P_t. \quad (5)$$

Here, N indicates the number of instances present in the test set and $|T|$ is total number of all *AP*s computed for each class.

Compliance accuracy of retail image with respect to company-provided planogram layout is calculated by the following equation, where $P_{matched}$ is the number of matched SKUs and P_{total} indicates the total number of SKUs:

$$\textbf{Compliance Accurancy} = 1 - \frac{\left(P_{matched} - P_{total}\right)}{P_{total}}. \quad (6)$$

4.2. Dataset Description. There is always a need for a huge amount of data for deep learning models. For evaluating effectiveness of the proposed approach, the dataset used in this study is provided by the industry partner which contains products of different categories and subcategories used for fine-grained recognition; the dataset contains 30,000 images, all collected manually from real diverse environment, that is, retails, departmental stores, marts and supermarkets, and shops with natural lightings through different mobile cameras.

The average resolution of the images was 1024×1024 with jpeg format. Images were captured from the distance of 1.5 to 5 feet from the front of the shelves. Each image contains multiple products. The testing set was also collected from real-life scenario through handheld devices. The dataset has a hierarchical structure containing a total of 12 main categories which cover diverse appearance, for example,, boxes, bottles, poach, and chiller, and contains 106 different fine-grained SKUs. Figure 9 shows the number of SKUs in each subcategory.

There is very minor difference in the packages of fine-grained categories. Rich annotations are provided to each product including the category, count, sizes, and flavors. 3000 images were labeled by using a labeling tool called LabelImg as presented in Figures 10(a) and 10(b). As each image contains multiple SKUs, almost 50 or more, high accuracy can be achieved during training. The percentages of training set and testing set were 80% and 20%, respectively. Training set and testing set contain 2400 and 600 images, respectively. The images of racks in the dataset were collected from 100 different types of racks which approximately contain six different levels. The annotation tool we used gave .txt file for each image. The text file contains class and location information in the form of class number and *x, y, w, h* coordinates.

4.3. Quantitative Evaluation. For evaluating Hyb-SMPC, the Amazon Web Service (AWS) instance called Elastic Compute Cloud (EC2) is used. In this study, for training process, Graphic Processing Unit (GPU) used is NVIDIA Tesla V100. At first, the training of the detectors was performed one by one on the GPU. In the proposed study, the Darknet-based framework is used for YOLO V4, whereas YOLO V5 and YOLOR are based on PyTorch-based framework.

Transfer learning is the concept of reusing the knowledge acquired from one specific task in another related newer task. This makes the learning process fast and enhances the performance of the deep learning models. Various models have been trained on challenging datasets which are then used for tackling related problems. In this work, the pretrained model used was "yolov4.conv.137" for YOLO V4. During training process, the training dataset was divided into small units called batches to perform learning of models. In this work, we used batch size of 64 and number of epochs is 72000. Input size of images was 512×512, with learning rate of 0.00261.

The training progress plot of the best category is illustrated in Figure 11. This plot helped us in monitoring the training process which is showing the "training accuracy." The details of average precision achieved by three different detectors trained for different categories are given in Table 4.

The highest average precision of 99% was achieved for the categories of coffee, milk modifier, and powder tea whitener. Furthermore, comparison of *mAP*, recall, and *F1-*

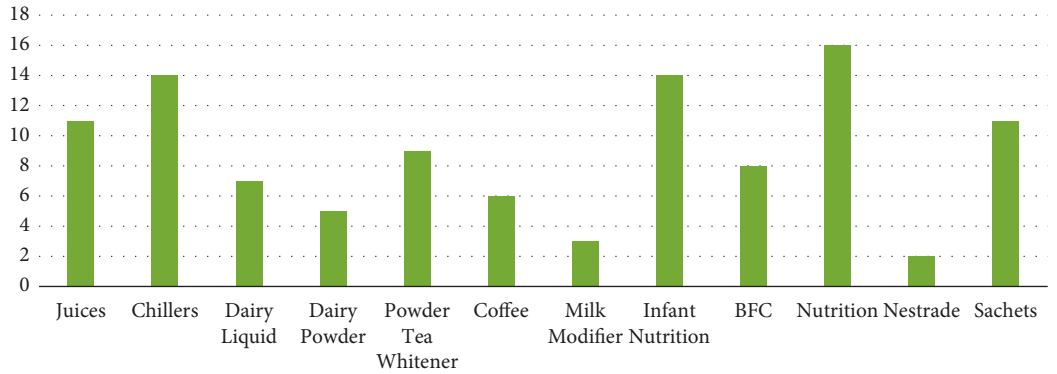
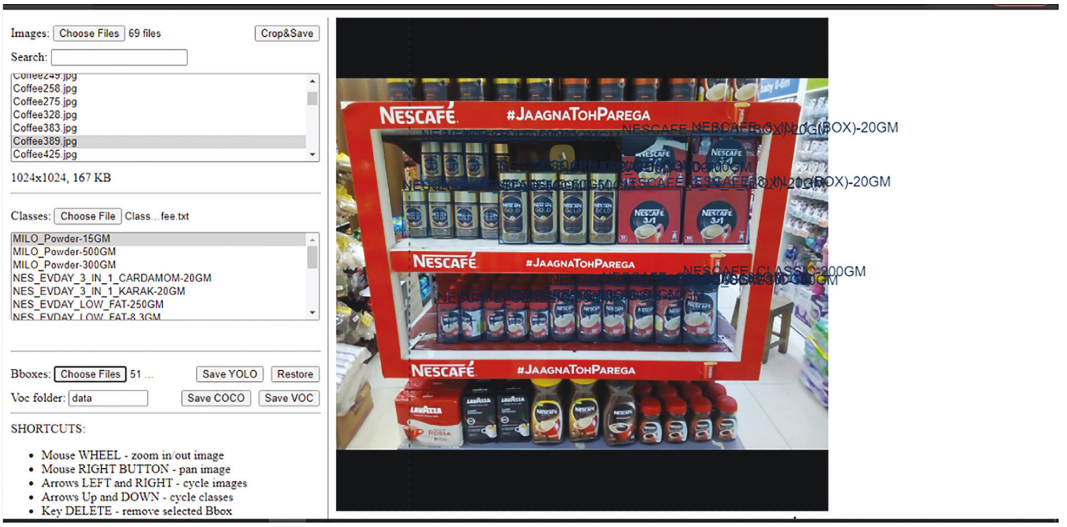
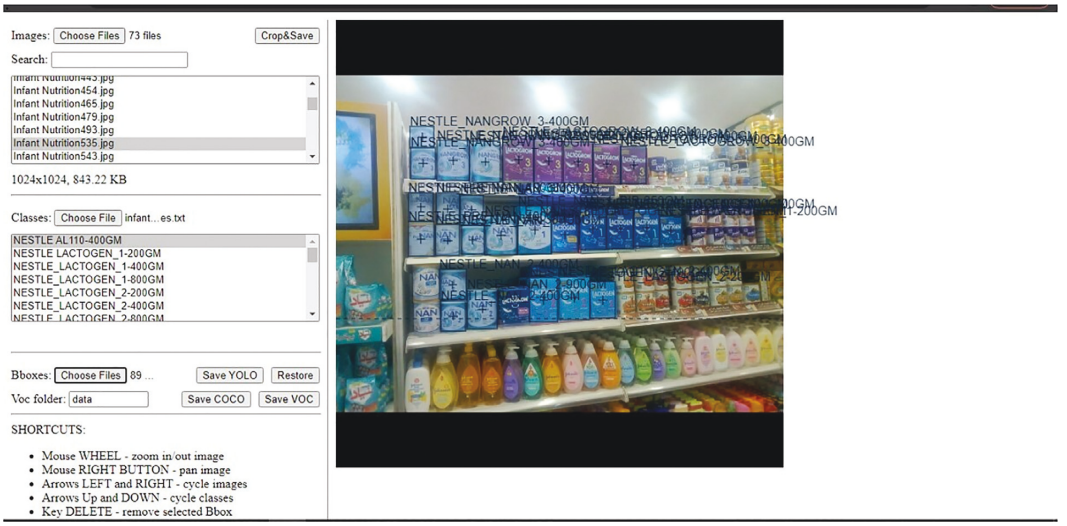


Figure 9: Number of SKUs in each category.



(a)



(b)

Figure 10: Labeled images. (a) Labeled image for coffee. (b) Labeled image for infant nutrition.

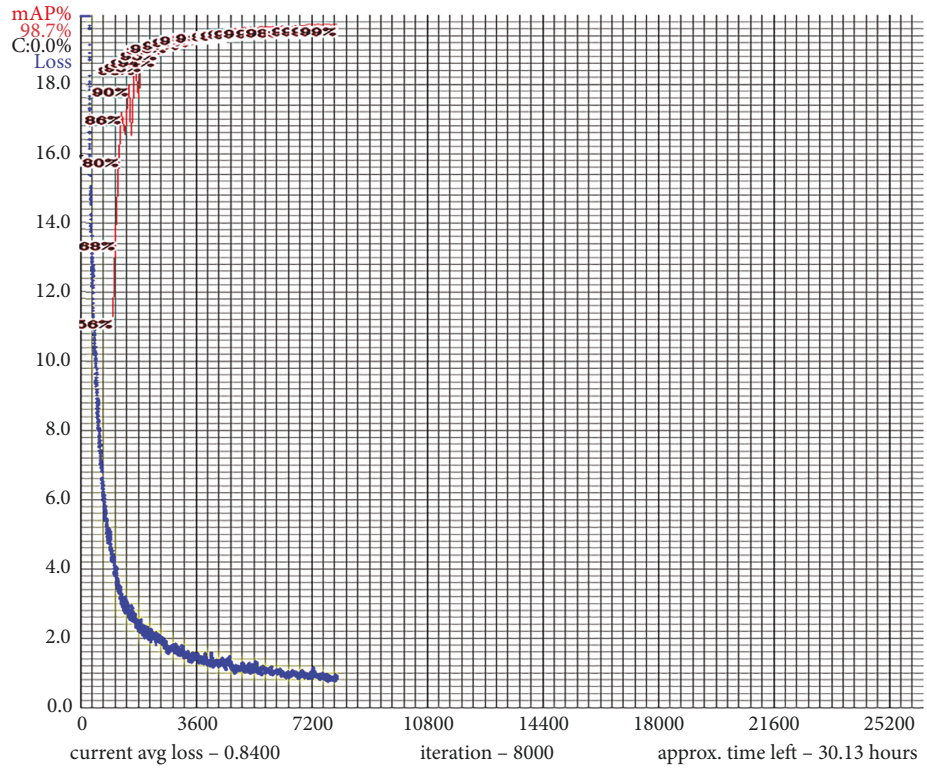


Figure 11: Training progress plot for categories of coffee, milk modifier, and PTW.

Table 4: Comparison of average precision achieved by three different detectors.

Main categories	YOLO V4	YOLO V5	YOLOR
		AP	
Juices	0.97	0.81	0.77
Chiller	0.97	0.82	0.79
Dairy liquid	0.965	0.826	0.89
Dairy powder	0.966	0.866	0.87
Coffee	0.987	0.85	0.85
Milk modifier	0.982	0.85	0.84
PTW	0.984	0.88	0.87
Infant nutrition	0.95	0.84	0.80
BFC	0.92	0.88	0.73
Nutrition	0.92	0.82	0.79
Nestrade	0.95	0.81	0.76
Sachets	0.95	0.81	0.78

score of three different detectors is provided in Table 5 and graphically demonstrated in Figure 12. The details of the planogram compliance accuracy achieved by Hyb-SMPC for different categories of the SKUs are provided in Table 6.

To evaluate the significance of Hyb-SMPC with the conventional methods, the test cases based on the size of the products are made. The effectiveness of the Hyb-SMPC is demonstrated in Table 7, which provides the comparison of the proposed approach with the conventional methods of [5, 22]. The results indicate that the Hyb-SMPC outperformed the conventional methods.

4.4. Qualitative Evaluation. The study also presents the qualitative evaluation of the proposed approach; for this purpose, the user's feedback is collected and analyzed. The users are divided into two groups; both groups provided their feedback by completing a survey which is incorporated in the annexure (included as separate file). We report on group 1 (retailer's group) as it is the most significant. The findings from both groups are presented below.

4.4.1. Retailer's Feedback. This group is comprised of professional retailers working in the domain of retailing, and the following are the summarized results:

(1) All the members were pleased to see new automated system.

(2) All the members felt content using the new system, checking reporting mechanism, and reviewing it.

Table 5: Comparison of *mAP, recall*, and *F1-score* of three different detectors.

Metrics	YOLO V4	YOLO V5	YOLOR
mAP	0.96	0.833	0.82
*F*1-score	0.95	0.822	0.801
Recall	0.898	0.827	0.81

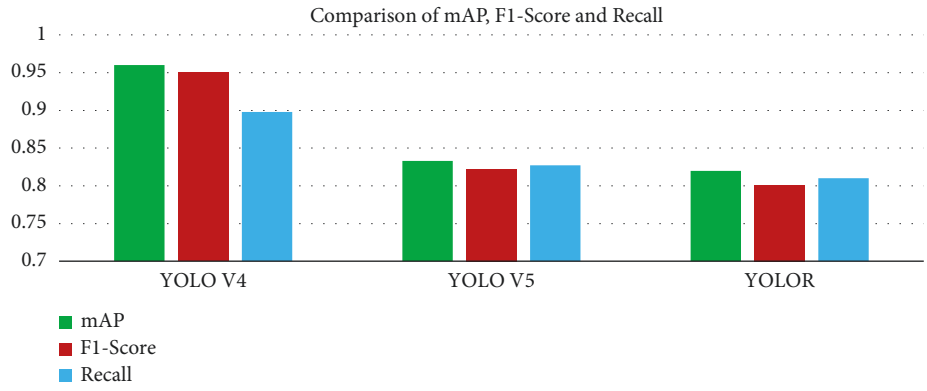


Figure 12: Success rate of three detectors.

Table 6: Compliance accuracy of Hyb-SMPC for different categories.

Categories	Compliance accuracy (%)
Juices	99.8
Chiller	99.6
Dairy liquid	98.1
Dairy powder	98.3
Coffee	99.7
Milk modifier	99.6
Powder tea whitener	99.65
Infant nutrition	97.5
BFC	96.5
Nutrition	96
Nestrade	97.43
Sachets	97.2

Table 7: Comparison of *compliance accuracy* of Hyb-SMPC with conventional method.

Product size	Recurring patterns (%)	Hyb-SMPC (%)
Small	95.32	99
Medium	90.61	97
Large	85.24	98

(3) Almost all members were enthusiastic about incorporating the new system to enhance their work efficiency.

Regarding the survey, statements S1 to S7 express positive statements for the approach. The responses obtained regarding all these statements were 4 or 5 (4: agree, 5: strongly agree) expect for S7 which mostly got the response of 4 (agree). To summarize, S1 obtained 100% response, S2 got 98%, S3 achieved 96%, S4 and S5 got 95%, and S6 achieved 97%, whereas S7 got 96%. Figure 13(a) represents level of user satisfaction regarding the aspects of system. A Wilcoxon signed-rank test was performed to determine whether there was a difference in the retailer's satisfaction level by using our approach compared to the previous manual technique. There was a statistically significant difference between the groups at 0.05 level; the p value equals 0.0156250; the test statistic Z equals −2.417559, which is not in the 95% region of acceptance: [−1.9600: 1.9600]. W = 0.0, is not in the 95% region of acceptance: [3.0000: 24.0000]. The observed standardized effect size, $Z/\sqrt{n}$, is large (0.86). That indicates that the retailers are quite satisfied with our approach.

Statements S8 to S14 represent negative statements regarding the presented approach. All the participants gave the score of 1 or 2 (1: strongly disagree, 2: disagree) except for S11 where *10.5%* responded with "agree," thus collectively demonstrating higher level of user satisfaction represented in Figure 13(b).

4.4.2. Customer's Feedback. This group consists of customers (both males and females) visiting the retail stores. The results obtained from the questionnaire provided to them indicate that the participants gave a score of 4 or 5 to the positive statements. S1 and S3 obtained *100%* response, whereas S2 obtained *89%*. The participants gave the score of 1 or 2 to the negative statements; S4 to S6 indicate negative statements. Only *2%* of participants were undecided about S4. Figure 14 shows the feedback of customers.

The results indicated that properly organized products increase the satisfaction level of the customers and let the customers visit the stores more often, which contributes to increasing the sales of the stores. Hence, the proposed approach can enhance the sales of stores to a significant level.

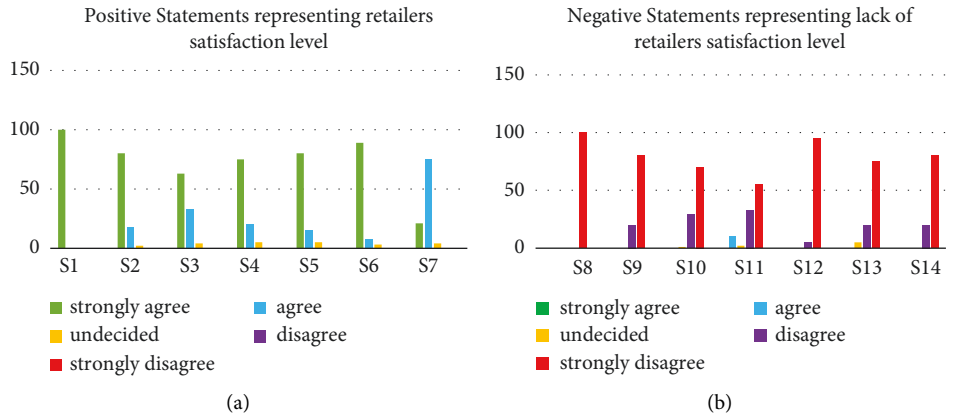


Figure 13: Retailer satisfaction level. (a) Positive statements. (b) Negative statements.

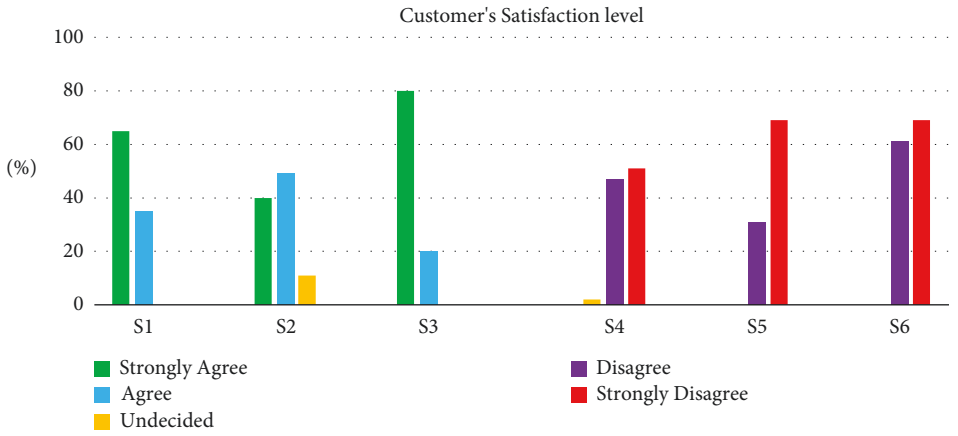


Figure 14: Customer's satisfaction level.

5. Conclusion

Effectively monitoring retail shelves and satisfying planogram are the two main factors that can boost sales of retail sector. The earlier studies conducted in this domain were either performing shelf monitoring or checking planogram compliance. As both activities are important, the proposed study presented a hybrid approach that deals with both activities. The study presented an approach to detect fine-grained retail products using deep learning and also verify the compliance of planogram. For the detection part, three one-stage detectors, YOLO V4, YOLO V5, and YOLOR, were trained on the dataset consisting of 30,000 retail images having 106 different SKUs belonging to 12 main categories. The use of one-stage detector makes the detection part fast. The best trained model performed efficiently on real retail environment and achieved accuracy up to 99%. The proposed method also checked planogram compliance by matching the provided planogram with the postprocessed images of retail and generated report indicating that the planogram is followed fully or partially or is not followed at all. There can be several extensions of the presented work. Some of the future considerations are described as follows:

(i) Augmenting Internet of things (IoT) to automate the manual process of capturing images by the personnel, instead the cameras mounted at different locations of store will capture the images and upload them on the servers for further processing.

(ii) Using strong quality assessment techniques to monitor the quality of captured images. In case of blurry, noisy, and distorted images, the system must not accept such images and ask the image-capturing entities to capture the images again. This technique will help improve accuracy.

(iii) Another extension of this work is to formulate a way to work with unlabeled data as manually labeling the SKUs in the images is a time-consuming and laborious task.

Data Availability

Data will be provided if required.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Supplementary Materials

The authors have supplied the description of dataset as supplementary materials. The dataset used is provided by the industry partner which contains products of 12 different categories (e.g., boxes, bottles, poach, chiller, etc.) and subcategories of 106 different fine-grained SKUs. The dataset contains 30,000 images manually collected from real diverse environment, that is, retails, departmental stores, marts and supermarkets, and shops with natural lightings through different mobile cameras. The average resolution of the images was 1024×1024 with jpeg format. Images were captured from the distance of 1.5 to 5 feet from the front of the shelves. (*Supplementary Materials*)

References

[1] B. Knezevic, S. Renko, N. Knego, and N. Knego, "Changes in retail industry in the Eu," *Business, Management and Education*, European Online Library, vol. 9, no. 1, pp. 34–49, 2011.

[2] Y. Wei, S. Tran, S. Xu, B. Kang, and M. Springer, "Deep learning for retail product recognition: challenges and techniques," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 8875910, 23 pages, 2020.

[3] A. Tonioni and L. Di Stefano, "Product recognition in store shelves as a sub-graph isomorphism problem," in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 682–693, LNCS, Catania, Italy, September 2017.

[4] M. Marder, S. Harary, A. Ribak, Y. Tzur, S. Alpert, and A. Tzadok, "Using image analytics to monitor retail store shelves," *IBM Journal of Research and Development*, vol. 59, no. 2/3, pp. 1–3, 2015.

[5] S. Liu and H. Tian, "Planogram compliance checking using recurring patterns," in *Proceedings of the 2015 IEEE International Symposium on Multimedia (ISM)*, Miami, FL, USA, December 2015.

[6] T. Elbers, *The Effects of In-Store Layout- and Shelf Designs on Consumer Behaviour*, 2016.

[7] D. Corsten and T. Gruen, "Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks," *International Journal of Retail & Distribution Management*, vol. 31, no. 12, pp. 605–617, 2003.

[8] X. Zhang and G. Wang, "Stud pose detection based on photometric stereo and lightweight YOLOv4," *Journal of Artificial Intelligence and Technology*, vol. 2, no. 1, pp. 32–37, 2021.

[9] A. Shabbir, A. Rasheed, A. Rasheed et al., "Detection of glaucoma using retinal fundus images: a comprehensive review," *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2033–2076, 2021.

[10] S. Karimi Jafarbigloo and H. Danyali, "Nuclear atypia grading in breast cancer histopathological images based on CNN feature extraction and LSTM classification," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 4, pp. 426–439, 2021.

[11] M. Merler, C. Galleguillos, and S. Belongie, "Recognizing groceries in situ using in vitro training data," in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 2007.

[12] B. Santra and D. P. Mukherjee, "A comprehensive survey on computer vision based approaches for automatic identification of products in retail store," *Image and Vision Computing*, vol. 86, pp. 45–63, 2019.

[13] K. Mikolajczyk and K. Mikolajczyk, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[14] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded up robust features," in *Proceedings of the Computer Vision - ECCV 2006*, pp. 404–417, Graz, Austria, May 2006.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] R. Moorthy, S. Behera, S. Verma, S. Bhargave, and P. Ramanathan, "Applying Image Processing for Detecting On-Shelf," in *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pp. 451–457, Kochi, India, August 2015.

[17] T. Mittal, B. Laasya, and J. Dinesh Babu, "A logo-based approach for recognising multiple products on a shelf," in *Proceedings of the SAI Intelligent Systems Conference (IntelliSys) 2016*, vol. 16, pp. 15–22, London, UK, September 2018.

[18] G. Varol and R. S. Kuzu, "Toward retail product recognition on grocery shelves," in *Proceedings of the Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, vol. 9443, pp. 1–7, Beijing, China, October 2014.

[19] A. Saran, E. Hassan, and A. K. Maurya, "Robust visual analysis for planogram compliance problem," in *Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pp. 576–579, Tokyo, Japan, May 2015.

[20] E. Frontoni, M. Contigiani, G. Ribighini, and I. Dii, "A Heuristic Approach to Evaluate Occurrences of Products for the Planogram Maintenance," in *Proceedings of the 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, Senigallia, Italy, September 2014.

[21] W. Geng, F. Han, J. Lin et al., "Fine-grained grocery product recognition by one-shot learning," in *Proceedings of the 26th ACM international conference on Multimedia*, vol. 2, pp. 1706–1714, Seoul, Republic of Korea, October 2018.

[22] S. Liu, W. Li, S. Davis, C. Ritz, and H. Tian, "Planogram Compliance Checking Based on Detection of Recurring Patterns," *Computer Vision and Pattern Recognition*, vol. 3, pp. 1–8, 2016.

[23] N. O. Mahony, S. Campbell, A. Carvalho et al., "Deep Learning vs. Traditional Computer Vision," *Cv*, 2019.

[24] M. A. Aslam, M. N. Salik, F. Chughtai, N. Ali, S. H. Dar, and T. Khalil, "Image classification based on mid-level feature fusion," in *Proceedings of the 2019 15th International Conference on Emerging Technologies (ICET)*, Peshawar, Pakistan, December 2019.

[25] "Detection and prediction of traffic accidents using deep learning techniques," *Angewandte Chemie International Edition*, vol. 6, no. 11, pp. 951-952, 2022.

[26] S. Fatima, N. Aiman Aslam, I. Tariq, and N. Ali, "Home security and automation based on internet of things: a comprehensive review," in *Proceedings of the IOP Conference Series: Materials Science and Engineering*, vol. 899, no. 1, Article ID 12011, Chennai, India, September 2020.

[27] Q. Zou, K. Xiong, Q. Fang, and B. Jiang, "Deep imitation reinforcement learning for self-driving by vision," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 4, pp. 493–503, 2021.

[28] N. A. Othman and I. Aydin, "A new IoT combined body detection of people by using computer vision for security application," in *Proceedings of the 2017 9th International*

Conference on Computational Intelligence and Communication Networks (CICN), vol. 2018, pp. 108–112, Girne, Northern Cyprus, September 2017.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016, pp. 779–788, Las Vegas, NV, USA, June 2016.

[30] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: scalable and efficient object detection," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Article ID 10778, Seattle, WA, USA, June 2020.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[32] P. Jund, N. Abdo, A. Eitel, and W. Burgard, *The Freiburg Groceries Dataset*, http://arxiv.org/abs/1611.05799, 2016.

[33] E. Goldman and J. Goldberger, "CRF with deep class embedding for large scale classification," *Computer Vision and Image Understanding*, vol. 191, pp. 1–11, 2019.

[34] K. Higa and K. Iwamoto, "Robust estimation of product amount on store shelves from a surveillance camera for improving on-shelf availability," in *Proceedings of the 2018 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, Krakow, Poland, October 2018.

[35] K. Higa and K. Iwamoto, "Robust shelf monitoring using supervised learning for improving on-shelf availability in retail stores," *Sensors*, vol. 19, no. 12, pp. 2722–12, 2019.

[36] L. Karlinsky, J. Shtok, Y. Tzur, and A. Tzadok, "Fine-grained recognition of thousands of object categories with single-example training," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 965–974, Honolulu, HI, USA, July 2017.

[37] H. Sun, J. Zhang, and T. Akashi, "TemplateFree: product detection on retail store shelves," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 15, no. 2, pp. 242–251, 2020.

[38] T. Chong, I. Bustan, and M. Wee, "Deep learning approach to planogram compliance in retail stores," *Semant. Sch.*, pp. 1–6, 2016.

[39] R. Girshick, J. Donahue, T. Darrell, J. Malik, U. C. Berkeley, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 5000, Columbus, OH, USA, June 2014.

[40] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[42] Y. Konishi, Y. Hanzawa, M. Kawade, and M. Hashimoto, "Fast 6D pose estimation from a monocular image using hierarchical pose trees," in *Proceedings of the Computer Vision - ECCV 2016*, vol. 1, pp. 398–413, Amsterdam, The Netherlands, October 2016.

[43] R. Yilmazer and D. Birant, "Shelf auditing based on image classification using semi-supervised deep learning to increase on-shelf availability in grocery stores," *Sensors*, vol. 21, no. 2, pp. 327–426, 2021.

[44] P. Wajire and E. Pune, *Image classification for retail*, 2020.

[45] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[46] H.-W. Zhang, L.-J. Zhang, P.-F. Li, and D. Gu, "Yarn-dyed fabric defect detection with YOLOV2 based on deep convolution neural networks," in *Proceedings of the 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, vol. 17, pp. 170–174, Enshi, China, May 2018.

[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, http://arxiv.org/abs/2004.10934.

[48] J. Solawetz, "YOLOv5 New Version - Improvements And Evaluation," 2020, https://blog.roboflow.com/yolov5-improvements-and-evaluation/.

[49] C. Wang, I. Yeh, and H. M. Liao, "You Only Learn One Representation: Unified Network for Multiple Tasks," pp. 1–11, 2021, https://arxiv.org/abs/2105.04206.

[50] Codebrainz, "Color-names," 2021, https://github.com/codebrainz/color-names/blob/master/output/colors.csv.

[51] H. Y. Ha, "Integrating Deep Learning with Correlation-Based Multimedia Semantic Concept Detection," 2015.