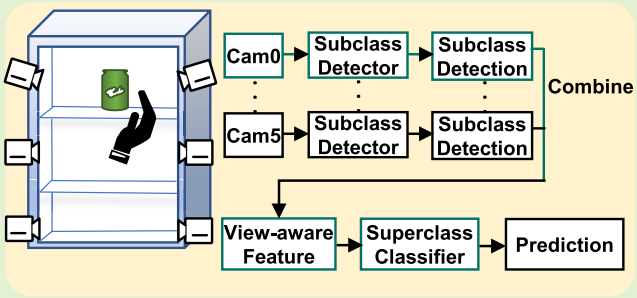
A Retail Object Classification Method Using Multiple Cameras for Vision-Based Unmanned Kiosks

Ji-Ye Jeon, *Graduate Student Member, IEEE*, Shin-Woo Kang, Hyuk-Jae Lee, *Member, IEEE*, and Jin-Sung Kim, *Member, IEEE*

***Abstract*—Several unmanned retail stores have been introduced with the development of sensors, wireless communication, and computer vision technologies. A vision-based kiosk that is only equipped with a vision sensor has significant advantages such as compactness and low implementation cost. Using convolutional neural network (CNN)-based object detectors, the kiosk recognizes an object when a customer picks up a product. In retail object recognition, the key challenge is the limited number of detections and high interclass similarity. In this study, these challenges are addressed by utilizing the "view-specific" feature of an object; specifically, an object class is divided into multiple "view-based" subclasses, and the object detectors are trained using these data. Further, the "view-aware feature" is defined by aggregating subclass detection results from multiple cameras. A superclass classifier predicts a superclass by utilizing an informative subclass detection result that distinguishes the target object from other similar-looking objects. To verify the effectiveness of the proposed approach, a prototype of the vision-based unmanned kiosk system is implemented. Experimental results indicate that the proposed method outperforms the conventional method, even on a state-of-the-art detection network. The dataset used in this study has been subsequently provided in the IEEE DataPort for reproducibility.**



***Index Terms*—Multicamera system, retail product recognition, unmanned retail, vending machine.**

I. Introduction

With the development of various sensors [1], [2], wireless communication [3], [4], and Internet of Things (IoT) technology [5], [6], many enterprises have integrated self-checkout systems in their businesses [7]. Recently, a "checkout-free" system that enables customers to simply walk out after selecting a product without interacting with any type of checkout system was introduced. Amazon Go [8] is a well-known example of fully automated checkout-free store. In Amazon Go, consumers purchase products in a grab-and-go style without any checkout process that involves tagging. These retail automations effectively reduce human labor, enhance customer shopping experiences, and increase convenience. However, these systems require numerous sensors and cameras and high computing power, and they are only applicable to large enterprises.

A vision-based unmanned kiosk that is equipped with only camera sensors can provide affordable options for reducing implementation costs. This kiosk is constructed with a display rack, in which an RGB camera is installed on the shelf. A convolutional neural network (CNN) detector is utilized to recognize a product that appears in the input image captured by the camera sensor. A key problem associated with product recognition is accuracy degradation caused by the limited number of detections and high interclass similarity. Retail products are infamous for their high intraclass appearance variations and high interclass appearance similarity [9]. For example, if a kiosk can only detect the flat top of a canned

Manuscript received 21 August 2022; revised 15 September 2022; accepted 22 September 2022. Date of publication 5 October 2022; date of current version 14 November 2022. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] (Variable-precision deep learning processor technology for high-speed multiple object tracking) under Grant 2020-0-01080 and in part by the MSIT, South Korea, under the Information Technology Research Center (ITRC) Support Program supervised by the IITP under Grant IITP-2022-2020-0-01461. The associate editor coordinating the review of this article and approving it for publication was Dr. Brajesh Kumar Kaushik. *(Corresponding author: Jin-Sung Kim.)*

Ji-Ye Jeon and Hyuk-Jae Lee are with the Electrical and Computer Engineering Department, Seoul National University, Seoul 08826, South Korea.

Shin-Woo Kang was with the Electrical and Computer Engineering Department, Seoul National University, Seoul 08826, South Korea. He is now with TMAX, Inc., Seoul 06210, South Korea.

Jin-Sung Kim is with the Electrical Engineering Department, Sun Moon University, Asan 31460, South Korea (e-mail: jinsungk@sunmoon.ac.kr).

Digital Object Identifier 10.1109/JSEN.2022.3210699

drink, it would be difficult to distinguish between two different drinks because appearances from various viewpoints will be similar for similar retail products. If the discriminative features of an object are not utilized for a detection network, the object can be incorrectly classified into a class that corresponds to a similar object.

In this study, the confusion caused by high interclass similarity is addressed using multiple cameras. Each camera has a CNN detector, and each detector conducts detection separately based on its own input image. Among the multiple detection results from various viewpoints, the classifier identifies an "informative" viewpoint (view-specific features) favorable to object identification and classifies the object by combining the detection results obtained from multiple detectors. Specifically, this study presents answers to the following questions.

1) How can we extract distinctive view-specific features from each camera sensors?
2) How can we utilize the informative appearance when detection results from multiple viewpoints are aggregated?

To answer the first question, we propose a "view-based" annotation method that defines multiple subclasses for a single object, considering the intraclass variations of the object. Subsequently, the detection network is trained with these subclasses to maximize view-specific saliency. During training, binary cross-entropy loss is used for enabling the detector to calculate the classwise probability. A view-specific feature is then defined as a vector of subclasswise probability.

To answer the second question, a view-aware feature is constructed by summing view-specific features, and a random forest (RF) superclass classifier is trained with the feature. RF has two advantages—it calculates the relative importance of given features and is robust to missing values. When multiple subclass detections are provided, RF identifies the "informative" viewpoint, where an object has a distinct appearance to distinguish it from other objects with similar appearances.

An approach similar to this is part-based detection [10], which is primarily used for the fine-grained classifications of various animals and artificial items [11], [12]. These methods annotated parts that typically appeared in the target object and extracted feature points from these parts. The extracted features were then aligned along a predefined order and input into the classifier. Despite their adequate performance, they cannot be applied to vision-based unmanned kiosks with multiple targets that do not have common parts. Furthermore, objects are often rotated during the purchase process, leading to additional computational overhead toward feature alignment. In addition, the aforementioned methods use datasets [13], [14], [15] consisting of on-the-shelf images or product-pack shots captured from a single camera. Classification utilizing a multicamera system has seldom been explored, and human action has not been considered.

The proposed classification method is summarized as follows. First, the newly proposed annotation method defines the subclasses of objects, which are then annotated using subclasses based on their appearances. The detection network is then trained using the subclass definition, and during inference, every implemented camera performs subclass detection. The final decision is made by aggregating the subclass detection results from multiple viewpoints. To evaluate the proposed algorithm, a prototype of a vision-based unmanned kiosk system is developed, along with a real-world retail product dataset, which comprises 142 420 images of shopping actions. The experimental results show that the proposed method outperforms the conventional method by 33.67% in terms of F1 score. In summary, this study makes the following contributions.

1) *Prototype of a Vision-Based Unmanned Kiosk:* The unmanned kiosk system equipped with multiple cameras is developed, and datasets are collected under this environment.
2) *View-Aware Classification Method:* A view-aware classification method along with a view-based annotation method, is proposed to resolve the interclass similarity of retail objects and to aggregate multiview detections.
3) *Real-World Dataset for a Vision-Based Unmanned Kiosk:* A real-world dataset capturing purchase process is devised with 142 420 images and two types of label sets. One label set has "view-based" labels on which the proposed annotation method is applied, and the labels in the second label set are annotated using a conventional method. For reproducibility, the datasets have been provided in the IEEE DataPort [16].

II. Related Work

A. Unmanned Retail Stores With Smart Checkout

The key characteristics of unmanned retail stores is a smart checkout system. Examples of the system include BingoBox's unstaffed convenience store, in which customers place products on a smart checkout counter that uses image recognition to calculate purchase prices. TaoCafé is another example that utilizes smart checkouts through facial recognition and a series of exit scanners. Amazon Go eliminated the checkout process, and customers can pick up products and walk out directly. This "Just Walk Out" system is enabled by built-in weight sensors and overhead cameras. Hundreds of cameras detect every customer and track them from the moment they enter the store until they leave; however, this real-time detection and tracking requires a high amount of computational power. It is worth noting that all smart checkout systems have been launched by large enterprises rather than small businesses.

This article proposes a vision-based unmanned kiosk as a compact version of an unmanned store; the kiosk detects retail products only using multiple cameras and classifies them using a simple machine learning algorithm with low computational cost. Additionally, no human detection such as face recognition is required because only one person can use the kiosk at a time; hence, possible privacy issues can be avoided when customers are detected.

B. Retail Object Recognition and Existing Datasets

Research on retail object recognition is highly relevant for the realization of real-world applications, and therefore, the real-world environments should be well reflected in the dataset. One study [15] proposed a method for enhancing

the localization performance of densely packed objects and developed a new benchmark (SKU-110K) containing images of supermarket shelves. Another study [17] recorded 36 videos capturing products within a shopping cart to test the smart self-checkout cart. In this study, a specific dataset that reflects the proposed kiosk system with multiple cameras is constructed. The established dataset includes various situations of the purchase process, such as frequent occlusions and viewpoint changes that can be caused by human actions and the use of multiple cameras. The dataset closest to our environment is [18]; however, the target retail product is limited to beverages, and only a single camera is used.

C. Utilizing Semantic Part Features

Several studies have been conducted to improve the performance of object recognition in severe environments such as frequent occlusion, fine-grained classification, and viewpoint changes. In their study, Girshick et al. [19] detected an object with handcrafted features from each part. Another study [10] further extended this study by annotating multiple parts of an object and building an object-part graphical model. Recently, an end-to-end network [20] was proposed, in which a network was trained to associate the local visual concepts with the semantic parts of the object. The network in [20] exhibited a significant improvement in performance under occlusion. Similarly, many studies have utilized the "part visual concept" in fine-grained classification tasks [11], [21], with [22] showing these approaches can be applied to retail product classification. In the previous studies, extracted part features were aggregated via concatenation and pooling for classification. Zhang et al. [11] concatenated part features with whole object features to obtain a pose-normalized representation. Zhang et al. [21] utilized an ROI-pooling layer to reorganize part features in a predefined order. Moreover, Srivastava [22] combined part features by average pooling. Although existing reports have highlighted their effectiveness in classifying objects observed from a "fixed" viewpoint, multiview object classification has seldom been explored. Furthermore, the target objects of unmanned kiosks do not have common parts, and therefore, ROI-pooling and re-ordering features is not applicable.

III. Proposed View-Aware Classification

A. Overview

Fig. 1 shows the entire process of the proposed method. First, multiple subclass detectors are distributed to each camera to detect the subclass defined by the proposed view-based annotation method. The subclass detection results are then utilized as a view-specific feature that represents the specific appearance of an object observed from each viewpoint. Subsequently, a view-aware feature is constructed by combining view-specific features from multiple cameras. A superclass classifier determines the final superclass of an instance using a view-aware feature. The details of view-aware detection, including the view-based annotation method and view-specific features, are presented in Section III-B. The view-aware classification is described in Section III-C. The specifications of the prototype unmanned kiosk are presented in Section IV, along with those of the image collection method.

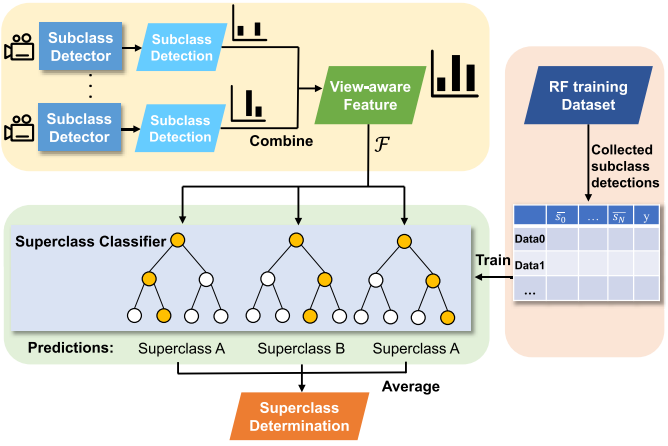


Fig. 1. Overview of the proposed method.

B. View-Aware Detection

1) View-Based Annotation: The goal of view-aware detection is to resolve the interclass similarity of the target objects by obtaining view-specific features. To obtain view-specific features, this study proposes a view-based annotation scheme. In the proposed view-based annotation, an object is divided into several subclasses considering its intraclass variations and is defined for the top, middle, and bottom of each object. The original category of an object is denoted as a superclass. Fig. 4 shows the superclasses and their subclasses in the proposed dataset. The proposed subclass definition helps improve recognition performance by separating distinct appearances from confusing appearances. For example, the easiest way to distinguish different canned beverages is to verify their product logo from the side view, thus serving as an informative differentiator between objects. By contrast, the top-view appearance will be a confusing subclass, as identifying the differences between the two canned beverages simply by observing at the flat top from the topview is difficult. The details of the annotation scheme and image collection are presented in Section IV.

2) View-Specific Feature: A view-specific feature is designed to represent the distinctive appearance observed from each viewpoint. To maximize view-specific saliency, a CNN is trained using a view-based annotated dataset; subsequently, the view-specific features are defined as vectors of subclass probabilities that are extracted from the CNN. The class confidence score indicates the probability of the class and how well the detected bounding box fits the object. When view-based annotation is applied, a single object is annotated with multiple subclass bounding boxes, which may overlap with each other. Therefore, binary cross-entropy is selected as a classification loss that enables multilabel classification. Each subclass has a probability that the predicted bounding box belongs to the subclass. We adopted YOLO [23], [24], a widely used real-time detection network that utilizes binary cross-entropy loss with a sigmoid function, as a subclass detector. When the bounding box i is predicted, its class confidence score of the jth subclass s_{ij} is defined by the

following equations:

$$s_{ij} = P\left(\text{Subclass}_j|\text{Object}\right) * P\left(\text{Object}\right)_i * \text{IOU}_i \quad (1)$$

$$= P\left(\text{Subclass}_j|\text{Object}\right) * \text{Objectness}\left(i\right). \quad (2)$$

Here, $P(\text{Object})_i$ is the probability that box i contains objects, and IOU_i refers to the intersection of union of box i. Objectness(i) is an objectness score obtained by multiplying $P(\text{Object})_i$ and IOU_i. Given that the sigmoid function is used to extract the classwise probability, each bounding box detection has a vector of class-specific confidence scores with values from 0 to 1. We only utilize features with scores above the threshold, θ. If the number of subclasses is N and the number of bounding boxes is B, then the view-specific feature C extracted from a single camera is expressed as follows:

$$\boldsymbol{C} = \sum_{i}^{B} \left(\bar{s_{i0}}, \bar{s_{i1}}, \ldots, \bar{s_{iN}}\right) \quad (3)$$

$$\bar{s_{ij}} = \begin{cases} 0, & 0 \le s_{ij} \le \theta \\ P\left(\text{Subclass}_j|\text{Object}\right), & \theta < s_{ij} \le 1. \end{cases} \quad (4)$$

Here, we use classwise probability $\bar{s_{ij}}$ because we empirically determined that classwise probability exhibits superior performance than that exhibited by class-specific confidence score, s_{ij}. The optimal value of θ is determined to be 0.25 with multiple experiments. It is worth noting that the constructed feature does not require geometric information such as the camera position or orientation of an object. Therefore, this feature is robust to sudden changes in pose of the object that may occur during the purchase process.

C. View-Aware Feature and Superclass Classifier

A superclass classifier determines the most probable superclass when multiple view-specific features are provided. To achieve this goal, a new feature, obtained by aggregating features from multiple viewpoints, is proposed. The descriptor is built using a view-aware feature $\mathcal{F}$, by extending the view-specific feature to a multicamera system. When there are N subclasses in a dataset and M cameras in the system, the view-aware feature $\mathcal{F}$ of an instance can be expressed as follows:

$$\mathcal{F} = \sum_{k}^{M} \boldsymbol{C}_k \quad (5)$$

$$= \sum_{k}^{M} \left(C_k\left(0\right), C_k\left(1\right), \ldots, C_k\left(N\right)\right) \quad (6)$$

$$= \left(\sum_{k}^{M} C_k\left(0\right), \sum_{k}^{M} C_k\left(1\right), \ldots, \sum_{k}^{M} C_k\left(N\right)\right). \quad (7)$$

Here, $C_k(j)$ refers to the jth element of the kth view-specific feature. The superclass classifier determines the final superclass by identifying the relative importance of given subclasses and the RF classifier [25], a widely used machine learning method, determines which subclass from $C_k(0)$ to $C_k(N)$ is "informative." The RF leverages the power of multiple decision tree. A decision tree is a treelike structure with multiple branches that determine the final output. A tree is constructed to maximize the information gain, which is the difference in entropy before and after the split is calculated. When a view-aware feature $\mathcal{F}$ is provided, the final prediction $\hat{y}$ is obtained by averaging the predictions from the trees.

In the proposed superclass classifier with the RF, an "informative subclass" has a more significant effect on superclass prediction. Given a particular prediction, $\hat{y}$, the importance of individual subclasses can be obtained by analyzing the decision path of the trees. This is accomplished by calculating subclass contributions using a previously reported method [26]. If a node is split by a subclass, the effect of the subclass on superclass prediction is defined as a change in superclass probabilities at the node. The superclass probability of a node is calculated by averaging the final prediction result of every training sample that flows into individual nodes. For instance, the superclass probability of the root node can be interpreted as a prior of the superclass because every training sample runs through the node. If the number of superclasses is P, the total contribution of subclass j for superclasses with view-aware feature $\mathcal{F}$, $\text{contr}(j, \mathcal{F})_t \in \mathbb{R}^P$, is obtained by summing the probability changes of superclasses in all nodes associated with subclass j along the decision path. Given that there are N subclasses, the prediction for a superclass c is defined as follows:

$$h_t\left(\mathcal{F}, c\right) = \text{bias}_t\left(c\right) + \sum_{j=1}^{N} \text{contr}\left(j, \mathcal{F}\right)_t\left(c\right). \quad (8)$$

Here, $\text{bias}_t(c)$ represents the cth element of the root node probability and $\text{contr}(j, \mathcal{F})_t(c)$ refers to the cth element of the contribution. Finally, the relative importance $\text{Importance}(\mathcal{F}, c)_j$ of subclass j for RF prediction is obtained by averaging all predictions of the individual trees

$$\text{Importance}\left(\mathcal{F}, c\right)_j = \frac{1}{T} \sum_{t=1}^{T} \text{contr}\left(j, \mathcal{F}\right)_t\left(c\right). \quad (9)$$

To train the RF classifier, a set of subclass detection results was collected using additional RF training dataset. The training feature of the classifier is defined as follows:

$$\mathcal{F}_{\text{train}} = \left(\sum_{k}^{M} C_k\left(0\right), \sum_{k}^{M} C_k\left(1\right), \ldots, \sum_{k}^{M} C_k\left(N\right), y\right). \quad (10)$$

Here, y denotes the ground truth superclass label. The specifications of the RF training dataset are described in Section IV.

It is worth noting that RF is trained to learn feature importance inherently by our view-aware annotation and view-aware feature, without any manual manipulations. We utilized the aforementioned equations to analyze the effect of the relative importance of each subclass in Section V.

IV. Datasets for the Vision-Based Unmanned Kiosk

A. Image Collection

Fig. 2 shows a prototype of the proposed vision-based unmanned kiosk, which consists of three levels of racks,

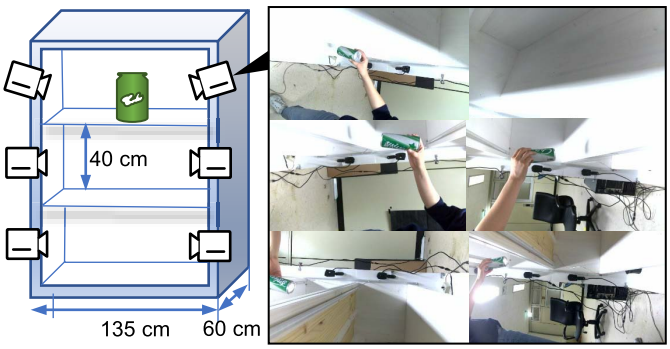


Fig. 2. Illustration of a prototype of the vision-based unmanned kiosk.

Fig. 3. Camera with the CMOS OV2710 image sensor installed in the vision-based unmanned kiosk.

each containing two cameras. Consequently, six cameras are present in the system. As shown in Fig. 3, high-definition USB cameras were installed at each end of the layer. Equipped with the 1/2.7" CMOS OV2710 image sensor, the camera can capture up to 60 frames/s at a resolution of 1280 × 720, with a 150° super-wide-angle lens. To mimic a real-world shopping environment, this study demonstrates scenarios in which six retail products are purchased. The entire process was recorded, and each frame was saved as 360 × 640 JPG images.

In the system, the classification is conducted on an instance-level, aggregating multiple detection results from multiple cameras. Therefore, multiple cameras should be synchronized to ensure the inputs are captured simultaneously. This can be guaranteed with multithread programming, where a thread computes the detection on each camera.

TABLE I
STATISTICS OF VIEW-BASED ANNOTATED DATASET AND CONVENTIONALLY ANNOTATED DATASET

train				valid			
View-based subclass	labels	Conventional superclass	labels	View-based subclass	labels	Conventional superclass	labels
C_{top}	736	Cider	1,477	C_{top}	731	Cider	2,028
C_{mid}	1,032			C_{mid}	2,017		
M_{top}	788			M_{top}	755		
M_{mid0}	757	Milk	2,352	M_{mid0}	709	Milk	1,982
M_{mid1}	816			M_{mid1}	539		
J_{top}	751	Jam	1,375	J_{top}	613	Jam	2,366
J_{mid}	1,037			J_{mid}	2,251		
T_{top}	778			T_{top}	749		
T_{mid}	956	Tuna	2,409	T_{mid}	761	Tuna	2,164
T_{bot}	807			T_{bot}	754		
B_{top}	736	Beer	2,013	B_{top}	592	Beer	1,477
B_{mid}	1,982			B_{mid}	1,468		
F_{top}	751	Freshener	2,697	F_{top}	992	Freshener	1,842
F_{mid}	1,937			F_{mid}	850		

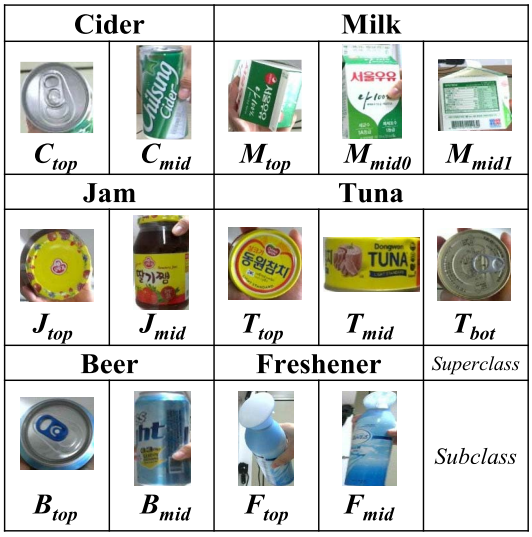


Fig. 4. Definition of superclasses and subclasses in the proposed view-based dataset.

B. Dataset Details and Statistics

1) Training Dataset: This study provides two types of training datasets. The first is a "view-based" dataset, to which the proposed view-based annotation is applied, and the second is a "conventional dataset," which is annotated in a typical manner using only superclasses. These two datasets share identical images; however, they have different label sets that are annotated with different schemes. Table I lists the statistics of the training and validation datasets. The dataset contains six retail products as superclasses, and the products were selected owing to their similar appearance from certain viewpoints. Fourteen subclasses are then defined for the six superclasses, and definitions of the superclasses and subclasses are shown in Fig. 4.

In the view-based dataset, the label data consists of bounding boxes and their subclass labels. The bounding box is drawn such that it completely covers the boundary of the object. From some viewpoints, a single product may be annotated using overlapping bounding boxes. The definition of a bounding box is not strictly limited because our objective is not to establish a precise box location for a single image. The total numbers of labels and images in the dataset are 48 323 and 48 364, respectively, including the validation set.

2) Test and RF Training Datasets: To evaluate the effectiveness of the proposed method, a multicamera dataset was constructed. In the dataset, a single instance consisted of six frames captured from six cameras. Unlike the training dataset, we denote only the superclass of an instance. The test dataset contains 10 306 instances. The number of instances for each superclass ranges from 1699 to 1780. The total number of images in the dataset is 61 836. The RF training dataset contains 5370 instances and 32 220 images, with 870–900 instances for each superclass. The statistics for the datasets are summarized in Table II. To reduce bias in the RF, we constructed an RF training dataset separate from the

TABLE II
STATISTICS OF TEST DATASET AND RF TRAINING DATASET

Superclass	test		RF training	
	images	instances	images	instances
Cider	10,680	1,780	5,400	900
Milk	10,200	1,700	5,400	900
Jam	10,200	1,700	5,400	900
Tuna	10,200	1,700	5,400	900
Beer	10,362	1,727	5,400	900
Freshener	10,194	1,699	5,220	870

TABLE III
OPTIMAL CONFIGURATIONS OF RF

	Tiny-YOLOv3	YOLOv3	YOLOv4
min_samples_split	2	2	2
min_samples_leaf	2	2	1
n_trees	600	100	100

training dataset by balancing the number of instances of each superclass.

V. EXPERIMENT

A. Experimental Setting

1) Detection Network: In the experiment, three types of real-time object detector YOLOv3 [23], Tiny-YOLOv3 [23], and YOLOv4 [24] are used as detection networks, and an identical training scheme is applied to train all networks. Networks are trained for 28 000 iterations with batch size 64. The learning rate is initialized to 0.0013, and decreased by 1/10 when the step number reaches 22 400 and 25 200 iterations. Conventional data augmentation is used with saturation 1.5, exposure 1.5, and hue 0.5.

2) Superclass Classifier: We construct a training dataset for RF with the subclass detection results collected from the RF training dataset in Section IV. Subsequently, we split the dataset into a sixfold cross-validation set. The optimal configurations of the RF is determined by varying the number of trees, minimum samples of leaf nodes, and minimum samples to split a node. The configurations identified for the different detection networks are listed in Table III.

B. Baselines

The proposed method utilizes a view-aware subclass detector and an RF superclass classifier, denoted by "View + RF." To verify the effectiveness of the proposed method, two baseline methods were compared.

The first baseline is a conventional approach that uses a conventional detector with the majority voting method, which is denoted as "Conv + Major." In this approach, the detector is trained using a conventionally annotated dataset with superclass. The final class decision is made by selecting the class that is most frequently detected. If the number of detections is identical, the class with the largest sum of the classwise probability is considered to be the final prediction.

The second baseline uses the view-aware subclass detector with the majority voting method, denoted as "View + Major." The detector is trained with the proposed view-based annotation method. The superclass of which the subclasses are most frequently detected is selected as the final prediction. Similar to the first baseline, the sum of the classwise probability was used when the number of subclass detections was the same.

C. Evaluation Metric for Multicamera Instance

For the multicamera instance evaluation, the Macro- and Micro-F1 scores presented in [27] are utilized. The Macro-F1 score ($F1_M$) is used to represent the overall performance of the method, while the Micro-F1 score ($F1_\mu$) represents the performance of each superclass. The Micro-F1 score ($F1_\mu$) is the harmonic mean of precision and recall of each class. The Micro-F1 score of superclass i can be calculated using the following equation:

$$\text{Precision}_{\mu_i} = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{11}$$

$$\text{Recall}_{\mu_i} = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{12}$$

$$\text{F1}_{\mu_i} = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \tag{13}$$

It is necessary to consider the definition of FN_i for an instance. As presented in Fig. 1, a superclass classifier determines a superclass using the detection results from multiple detectors. If every detector fails to recognize a product in an instance, the classifier cannot yield a prediction. Thus, this **missed case** must to be considered.

For every instance x in the dataset, we denote ground truth of x by $y(x)$, prediction of x by $\hat{y}(x)$ and the number of instances by l. The true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) of a superclass $i \in (0, 1, 2, \ldots, K-1)$ are defined as follows:

$$\text{TP}_{\mu_i} = \sum_{x:\hat{y}(x)=i}^{l} I\left(y\left(x\right) = i\right) \tag{14}$$

(15)

$$\text{FP}_{\mu_i} = \sum_{x:\hat{y}(x)=i}^{l} I\left(y\left(x\right) \neq i\right) \tag{16}$$

$$\text{TN}_{\mu_i} = \sum_{i \neq j}^{K} \text{TP}_{\mu_j} \tag{17}$$

$$\begin{aligned}\text{FN}_{\mu_i} &= \text{FN}_{w\,\mu_i} + \text{FN}_{m\,\mu_i} \\ &= \sum_{j \neq i}^{K-1} \sum_{x:\hat{y}(x)=j}^{l} I\left(y\left(x\right) = i\right) + \sum_{x:y(x)=y}^{l} \cdot 1 - \text{TP}_{\mu_i}.\end{aligned} \tag{18}$$

Here, $I()$ is an indicator function that returns one when the equation is true. FN_w refers to the FNs caused by incorrect predictions, and FN_m refers to the **missed case**. FN_m is calculated by subtracting TP from the number of true instances of a class.

Subsequently, the Macro-F1 score ($F1_M$) is computed as the arithmetic mean of the Micro-F1 scores ($F1_{\mu_i}$) of each class,

TABLE IV
MACRO-PERFORMANCE OF THE THREE METHODS

Method	Annotation	Superclass Classifier	Detection Network	$Precision_M$	$Recall_M$	$F1_M$
Conv+Major	Conventional	Majority Voting	Tiny-YOLOv3	0.824	0.766	0.774
			YOLOv3	0.822	0.682	0.686
			YOLOv4	0.872	0.760	0.765
View+Major	View-based	Majority Voting	Tiny-YOLOv3	0.924	0.860	0.889
			YOLOv3	0.894	0.832	0.849
			YOLOv4	0.918	0.884	0.895
View+RF	View-based	Random Forest Classifier	Tiny-YOLOv3	**0.944**	**0.880**	**0.908**
			YOLOv3	**0.937**	**0.901**	**0.917**
			YOLOv4	**0.962**	**0.943**	**0.951**

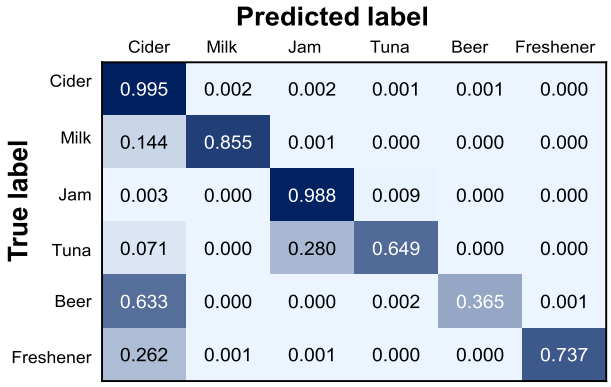


Fig. 5. Confusion matrix of Conv + Major applied on YOLOv4.

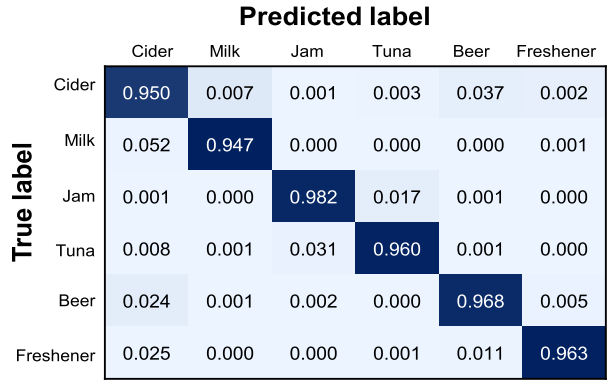


Fig. 6. Confusion matrix of View + RF applied on YOLOv4.

as follows:

$$\text{Precision}_M = \frac{\sum_{i=0}^{K-1} \text{Precision}_{\mu_i}}{K} \tag{19}$$

$$\text{Recall}_M = \frac{\sum_{i=0}^{K-1} \text{Recall}_{\mu_i}}{K} \tag{20}$$

$$\text{F1}_M = \frac{\sum_{i=0}^{K-1} \text{F1}_{\mu_i}}{K}. \tag{21}$$

D. Experimental Results

Table IV summarizes the experimental results of the three methods applied to the test set. The proposed superclass classifier with a view-aware detector (View + RF) outperforms all other baselines in every detection network. Compared to majority voting classification with the conventionally trained detector (Conv + Major), View + RF improves the accuracy on Tiny-YOLOv3, YOLOv3, and YOLOv4 by **17.31%, 33.67%, and 18.69%,** respectively.

Figs. 5 and 6 present the confusion matrixes of the conventional and the proposed methods, respectively. Figs. 5 and 6 show that the confusion caused by interclass similarity is significantly alleviated by the proposed method.

It is worth noting that the view-based annotation method and RF play an important role in the performance improvement. Table IV also shows that applying only the view-based annotation can improve the performance for all detection network. In YOLOv4, the view-aware detector with the majority voting method has **0.046 (+5.28%), 0.124 (+16.32%), and 0.130 (+16.96%)** points higher precision, recall, and F1 scores, respectively, compared to those of the conventionally trained detector with the majority voting method. Likewise, using RF with a view-aware detector improves the precision, recall, and F1 scores by **0.044 (+4.79%), 0.059 (+6.67%), and 0.056 (+6.26%)** points, respectively, compared with those obtained with majority voting. The effects of the annotation method and superclass classifier are further analyzed in Sections V-E and V-F.

E. Effect of the View-Based Annotation Method

First, we analyze the impact of the view-based annotation method on a **single camera object detection network**. Specifically, we compare the numbers of FP and TP of the proposed view-aware detector and conventionally trained detectors. Each detector is evaluated using the corresponding validation set described in Section IV. In Table V, we summarize the results of the evaluations performed on Tiny-YOLOv3, YOLOv3, and YOLOv4. As shown, view-aware detector reduces the FP by **33.77%, 58.61%, and 45.90%**, respectively, compared with the conventional detector. In the proposed view-based dataset, the intraclass variance is reduced as a single object is divided into multiple subclasses, which leads to a reduction in FPs. However, simultaneously, the intraclass bias increases, which leads to a considerable increase in the number of FNs. Overall, these single-view results show that the detection network trained with the view-based dataset has high precision but relatively low recall, which is desirable in a multicamera system because high precision ensures that each detection

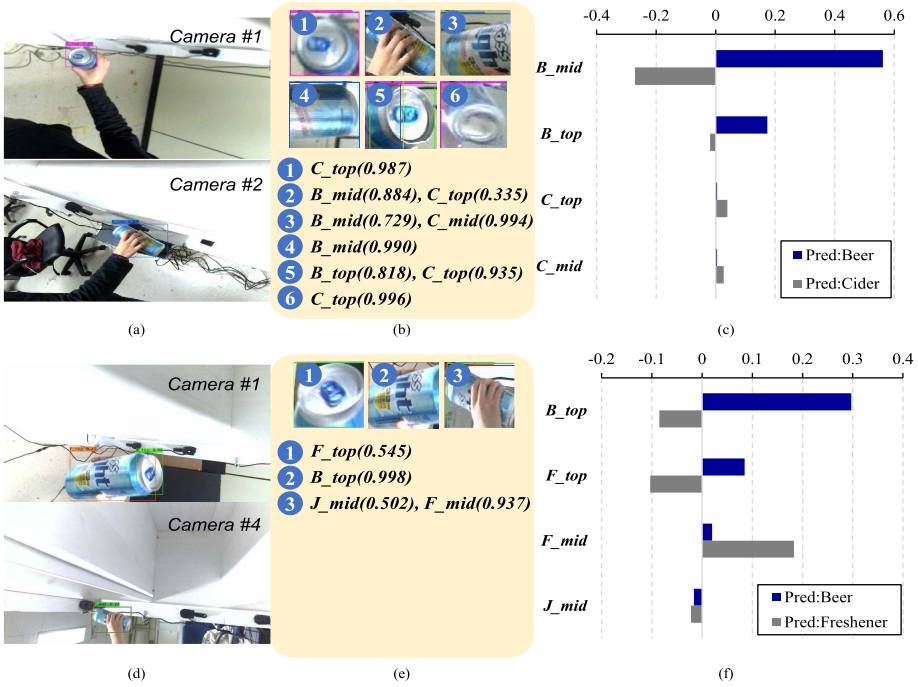


Fig. 7. Examples of instances that have been misclassified in View + Major and then corrected using View+RF. (a) Captured input frames. (b) Subclass detection results—each numbered circle denotes a detected bounding box and the classwise probability of each bounding box is shown in parentheses. (c) Subclass importance for the true prediction (Beer) and false prediction (Cider). (d) Captured input frames. (e) Subclass detection results. (f) Subclass importance for the true prediction (Beer) and false prediction (Freshener).

TABLE V
VARIATION IN THE DETECTION RESULTS FOR A SINGLE IMAGE WHEN THE VIEW-BASED ANNOTATION METHOD IS APPLIED TO THE DETECTION NETWORKS

Detection Network	Statistics	Conventional	View-based	Variation Rate (%)
Tiny-YOLOv3	# of FP	4,143	2,744	**-33.77**
	# of TP	7,596	6,981	-8.10
	# of FN	4,061	6,800	+67.45
	mAP@50 (%)	68.73	69.16	+0.62
YOLOv3	# of FP	6,656	2,755	**-58.61**
	# of TP	8,100	7,785	-3.89
	# of FN	3,557	5,966	+67.72
	mAP@50 (%)	70.58	69.16	-2.01
YOLOv4	# of FP	3,192	1,727	**-45.90**
	# of TP	10,426	10,635	+2.00
	# of FN	1,231	3,146	+155.56
	mAP@50 (%)	87.81	89.69	+2.14

TABLE VI
MICRO-PERFORMANCE OF "CONV + MAJOR" METHOD APPLIED ON YOLOV4

Superclass	TP	FP	FN_w	FN_m	Precsision	Recall	F1
Cider	1,741	**1,907**	9	24	0.478	0.981	0.673
Milk	1,441	4	245	14	0.997	0.847	0.916
Jam	1,673	**480**	20	7	0.777	0.984	0.868
Tuna	1,100	20	**595**	5	0.982	0.647	0.780
Beer	630	1	**1,097**	0	0.998	0.364	0.534
Freshener	1251	1	447	1	0.999	0.736	0.848
Sum/Avg.	7,842	2,413	2,413	**51**	0.872	0.760	0.765

result is sufficiently distinct; thus, it can serve to represent a specific viewpoint.

We now present the numerical results for the effect of the annotation methods in **multicamera classification** on the test dataset in Tables VI and VII. The models in Tables VI and VII are the conventional and view-aware detectors, respectively, and the superclass is determined using the majority voting method in both methods. The method with the view-aware detector outperforms that with the conventional detector in terms of the Micro-F1 score. Table VII shows that the view-aware detector reduces a large number of FPs and FN_ws for confusing pairs such as (Cider-Beer) and (Jam-Tuna). Consequently, the total number of TPs is significantly increased by **16.13%**, and the numbers of FP and FN_w are decreased by **59.10%**. These results demonstrate that when combined with a multicamera system, the proposed view-based annotation alleviates the classification error caused by interclass similarity. Tables VII and VIII show that View + Major misses the target object (FN_m) more frequently than Conv + Major, which stems from the precision-recall trade-off. This problem, which can be addressed by utilizing temporal information such as object tracking, can be explored in future work.

F. Effect of RF Classification

To validate the effectiveness of the proposed superclass classifier, the performance of the proposed RF is compared with that of the majority voting method. The same view-aware detection network is used in this experiment. Tables VII and VIII show that View + RF exhibits performance superior to that exhibited by View + Major across

TABLE VII
MICRO-PERFORMANCE OF "VIEW + MAJOR" METHOD APPLIED ON YOLOV4

Superclass	TP	FP	FN_w	FN_m	Precsision	Recall	F1
Cider	1,519	**623**	107	154	0.709	0.853	0.775
Milk	1,635	75	29	36	0.956	0.962	0.959
Jam	1,673	**212**	19	8	0.888	0.984	0.933
Tuna	1,373	10	**322**	5	0.993	0.808	0.891
Beer	1,287	2	**434**	6	0.999	0.745	0.853
Freshener	1,620	65	76	3	0.961	0.953	0.957
Sum/Avg.	9,107	987	987	**212**	0.918	0.884	0.895

TABLE VIII
MICRO-PERFORMANCE OF "VIEW + RF" METHOD APPLIED ON YOLOV4

Superclass	TP	FP	FN_w	FN_m	Precsision	Recall	F1
Cider	1,436	**75**	190	154	0.950	0.807	0.873
Milk	1,651	92	13	36	0.947	0.947	0.959
Jam	1,634	**30**	58	8	0.982	0.961	0.971
Tuna	1,662	70	**33**	5	0.960	0.978	0.969
Beer	1,642	54	**79**	6	0.968	0.951	0.959
Freshener	1,683	65	13	3	0.963	0.991	0.977
Sum/Avg.	9,708	386	386	212	0.962	0.943	0.951

all superclasses in terms of the Micro-F1 score. The subclass detection results for these two methods are identical; thus, the number of FN_m is the same, which implies that the improvement is solely achieved by correcting false predictions (FP, FN_w) in the proposed RF.

Fig. 7 introduces two examples in which Beer is misclassified by View + Major, whereas the proposed View + RF classifies it correctly. Fig. 7(a) shows two captured images of Beer, and Fig. 7(b) shows six images from the six cameras and the detection results with classwise probability. In Fig. 7, a view-aware detector confuses the subclass of Beer with that of Cider because they appear similar in the top view. The number of subclasses of Cider is larger than that of Beer, and thus, Cider is a superclass in the View + Major method. Fig. 7(c) shows the importance of each subclass for each superclass using (9). The subclass importance for the true prediction (Beer) is illustrated by the blue bar in the graph. By contrast, the subclass importance for the confusing class (Cider) is represented by the gray bar. Fig. 7(c) shows that B_{mid} contributes the most to the correction among the subclasses, and thus, the superclass is determined as Beer in View + RF. This result is in line with the initial hypothesis that two different canned beverages are confused by the similar appearance of the top view (B_{top}), and they can be distinguished by utilizing a discriminative side view (B_{mid}). In this case, B_{top} is an uninformative view and B_{mid} is an informative viewpoint.

Fig. 7(d)–(f) show another example in which Beer is misclassified as Freshener because the number of Freshener subclasses is larger than that of Beer. In this instance, the side view of Freshener is very similar to that of Beer, and the top view of Beer (B_{top}) is an informative view. Fig. 7(f) shows that this confusion is resolved with B_{top}, which contributes the most to the correct prediction of Beer in View + RF.

The two examples in Fig. 7 indicate that the informative subclasses are not fixed; however, they change with respect to the composition of the subclass detection results. In this study, view-based annotation divides an object into several subclasses considering intraclass variation, and the proposed superclass classifier of RF is trained to utilize the informative subclasses adaptively to resolve interclass similarity.

VI. CONCLUSION

This article proposed a retail product classification method for a vision-based unmanned kiosk system. To resolve the interclass similarity and limited number of detections, a multicamera system with view-aware classification method was introduced. The proposed method includes a view-based annotation method that enables a CNN detector to extract view-specific features with advantages in multicamera systems. The superclass classifier predicts the category of an object by collecting view-specific features, considering the relative importance of each viewpoint. For evaluation, a real-world dataset was constructed and released to the public. The experimental results demonstrated the effectiveness of the proposed method on various types of CNN detectors. Currently, the proposed system performs classification for instant input images. This method can be improved by utilizing temporal information and applying object tracking and can be explored in future research. Attention-based techniques are another promising approach for accuracy improvement, particularly when attention is used to identify relationships between regional features or the importance of specific features.

REFERENCES

[1] D. de Donno, L. Catarinucci, and L. Tarricone, "A battery-assisted sensor-enhanced RFID tag enabling heterogeneous wireless sensor networks," *IEEE Sensors J.*, vol. 14, no. 4, pp. 1048–1055, Apr. 2014.
[2] Q. Yuan, Z. Liu, F. Yang, and T. Ma, "Intelligent shopping cart design based on the multi-sensor information fusion technology and vision servo technology," *IEEE Sensors J.*, vol. 21, no. 22, pp. 26033–26041, Nov. 2021.
[3] N. Kumar and D. P. Vidyarthi, "A green routing algorithm for IoT-enabled software defined wireless sensor network," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9449–9460, Nov. 2018.
[4] Y. Zhou, W. Xiang, and G. Wang, "Frame loss concealment for multiview video transmission over wireless multimedia sensor networks," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1892–1901, Mar. 2015.
[5] F. Alawad and F. A. Kraemer, "Value of information in wireless sensor network applications and the IoT: A review," *IEEE Sensors J.*, vol. 22, no. 10, pp. 9228–9245, May 2022.
[6] T. Islam, S. C. Mukhopadhyay, and N. K. Suryadevara, "Smart sensors and Internet of Things: A postgraduate paper," *IEEE Sensors J.*, vol. 17, no. 3, pp. 577–584, Feb. 2017.
[7] Y. Wei, S. Tran, S. Xu, B. Kang, and M. Springer, "Deep learning for retail product recognition: Challenges and techniques," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–23, Nov. 2020.
[8] Amazon.Com. (2022). *Amazon Go*. Accessed: Jun. 11, 2022. [Online]. Available: https://www.amazon.com/b?ie=UTF8&node=16008589011
[9] I. Baz, E. Yoruk, and M. Cetin, "Context-aware hybrid classification system for fine-grained retail product recognition," in *Proc. IEEE 12th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jul. 2016, pp. 1–5.
[10] J. Zhu, X. Chen, and A. L. Yuille, "DeePM: A deep part-based model for object detection and semantic part localization," 2015, *arXiv:1511.07131*.
[11] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 834–849.

[12] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 26–33.
[13] M. Merler, C. Galleguillos, and S. Belongie, "Recognizing groceries in situ using in vitro training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[14] P. Jund, N. Abdo, A. Eitel, and W. Burgard, "The freiburg groceries dataset," 2016, *arXiv:1611.05799*.
[15] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5227–5236.
[16] J.-Y. Jeon and S.-W. Kang. (2021). *The Dataset for a Vision-Based Unmanned Kiosk*. [Online]. Available: https://ieee-dataport.org/documents/dataset-vision-based-unmanned-kiosk
[17] H.-C. Chi, M. A. Sarwar, Y.-A. Daraghmi, K.-W. Lin, T.-U. Ik, and Y.-L. Li, "Smart self-checkout carts based on deep learning for shopping activity recognition," in *Proc. 21st Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Sep. 2020, pp. 185–190.
[18] H. Zhang, D. Li, Y. Ji, H. Zhou, W. Wu, and K. Liu, "Toward new retail: A benchmark dataset for smart unmanned vending machines," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7722–7731, Dec. 2020.
[19] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 437–446.
[20] Z. Zhang, C. Xie, J. Wang, L. Xie, and A. L. Yuille, "DeepVoting: A robust and explainable deep network for semantic part detection under partial occlusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1372–1380.
[21] H. Zhang et al., "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.
[22] M. M. Srivastava, "Bag of tricks for retail product image classification," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2020, pp. 71–82.
[23] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[25] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1, Aug. 1995, pp. 278–282.
[26] A. Saabas. (2014). *Interpreting Random Forests*. Accessed: Jun. 11, 2022. [Online]. Available: http://blog.datadive.net/interpreting-random-forests/
[27] J. Opitz and S. Burst, "Macro f1 and macro f1," 2019, *arXiv:1911.03347*.

Ji-Ye Jeon (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2017, where she is currently pursuing the integrated M.S. and Ph.D. degrees in electrical and computer engineering.

Her research interests include image processing applications and automated machine learning.

Shin-Woo Kang received the B.S. degree in electrical and computer engineering from Han Yang University, Seoul, South Korea, in 2018, and the M.S. degree in electrical and computer engineering from Seoul National University, Seoul, in 2021.

In 2021, he joined TMAX, Inc., Seoul, as a Researcher. His current research interests include image processing applications.

Hyuk-Jae Lee (Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 1996.

From 1996 to 1998, he was with the faculty of the Department of Computer Science, Louisiana Tech University, Ruston, LS, USA. From 1998 to 2001, he was with the Server and Workstation Chipset Division, Intel Corporation, Hillsboro, OR, USA, as a Senior Component Design Engineer. In 2001, he joined the School of Electrical Engineering and Computer Science, Seoul National University, where he is currently a Professor. He is the Founder of Mamurian Design, Inc., Seoul, a fabless SoC design house for multimedia applications. His research interests include computer architecture and SoC design for multimedia applications.

Jin-Sung Kim (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1996, 1998, and 2009, respectively.

From 1998 to 2004 and from 2009 to 2010, he was with Samsung SDI Ltd., Cheonan, South Korea, as a Senior Researcher. From 2010 to 2011, he was a Postdoctoral Researcher with Seoul National University. In 2011, he joined the Department of Electronic Engineering, Sun Moon University, Asan, South Korea, where he is currently an Associate Professor. His current research interests include algorithm for video compression and computer vision.