Journal of Multimedia Information System VOL. 10, NO. 1, March 2023 (pp. 53-58): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2023.10.1.53

Lifelong Learning Architecture of Video Surveillance System

Taewan Kim[1]

Abstract

The learning capacity of general deep learning models for object detection would not be large enough to represent real-world scene dynamics, and thus such models would be weak to `unseen' data due to environmental changes. To address this issue, online or active learning methods use data samples obtained in new environments, where the new samples collected from false and/or miss detection cases are used to re-train the original model to enhance detection precision. However, it is inevitably degraded over time due to the catastrophic forgetting problem, that is a well-known intrinsic problem of current deep learning technologies. In this study, we propose a cutting-edge end-to-end system architecture to continuously improve the accuracy of the video analytic algorithms such as object detection with less accuracy degradation, by utilizing a hybrid combination of intelligence both the front-end and back-end systems. We use an iterative process where the current model is self-evolving using new incoming data as part of an ongoing adaptation process. We carried out several experiments of person detection in surveillance videos with various challenging environmental changes and showed the high precision and adaptability of our new architecture while it can be practically implemented at a low cost.

Key Words: Intelligent Video Analytics, Video Surveillance, Self-Adaptation, Person Detection.

I. INTRODUCTION

Recent advance in deep learning technologies has led to an explosion of several intelligent vision applications such as autonomous vehicles [1], computer-aided diagnosis [2] and video surveillance [3].

In particular for video surveillance, visual intelligence has been applied for helping and optimizing the manual decision making process by human vision as form of automatic scene understanding such as object detection with tracking, face recognition and abnormal behavior.

The visual intelligence is generally embedded in a machine learning (ML) model after training process with abundant data samples. Many researchers have validated their own ML models with a remarkable performance gain in their lab tests under a finite test sample set. But practically, after the model has been deployed in a video surveillance system, it would be quite frequent that unexpected data is fed into the system during 24/7 operations. The model performance normally tends to get worse or to converge at a low accuracy since the model would be naturally insufficient to understand all scene dynamics in real world.

The unseen data problem usually gets more severe for large-scale video surveillance such as video surveillance as a service, where the ML models are equipped on directly front-end edge devices and/or on back-end inference services over cloud. Under normal circumstances in practical applications, single ML model designed for a certain functionality such as intrusion detection is universally applied over all cameras and any form of customization in the ML model is unlikely to occur in real systems except for some limited parameters such as confidence threshold.

Many studies have attempted to elucidate how it may be resolved with data repurposing [4], domain adaptation [5] and novel training methods [6]. A transfer learning framework was proposed to convert a general pedestrian detector to a particular domain with a minimal annotation need [7]. In order to decrease the false negatives and false positives, authors in [8] suggested an unsupervised domain adaptation technique for a single-stage object detector employing weak self-training and adversarial regularization. Moreover, authors in [9] studied a theoretical framework and architecture for designing lifelong learning.

Proper training sources from false alarms and/or missed cases can overcome this limitation. However, they are infrequently accessible because of privacy legal problems for personal individual information, and unpredictability of new incoming data, as depicted in Fig. 1. Moreover, trying to update new data to the existing model requires modification of its weights, so that it contradicts the prior under-

Manuscript received February 27, 2023; Revised March 10, 2023; Accepted March 14, 2023 (ID No. JMI-23M-02-005)

Corresponding Author (*): Taewan Kim, +82-2-940-4751, kimtwan21@dongduk.ac.kr.
[1]Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, Korea, kimtwan21@dongduk.ac.kr

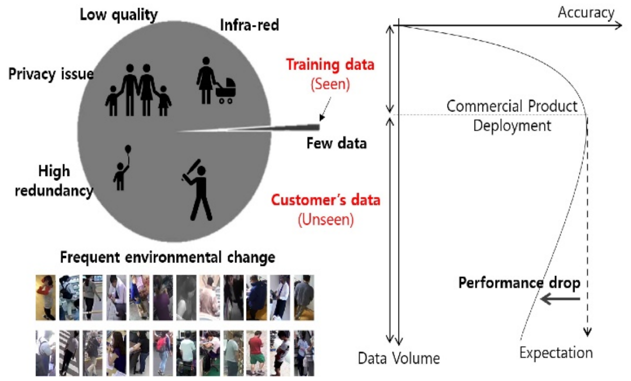

Fig. 1. Several challenges and characteristics of video surveillance.

standing significantly, yielding the issue know as catastrophic forgetting.

In this work, we have developed a novel end-to-end video surveillance architecture to adapt the ML model to the individual camera environment. detector in video surveillance using Multi-Scale ResBlock (MSRB). Then, a novel domain adaptation technique is employed in an intelligent back-end system to switch from a generic model to an individual-specific model by accurately understanding the space and context information.

To the best of our knowledge, the paper is the first research into a lifelong video surveillance system with a real-world commercial platform, driven mainly by the near flawless accuracy and scalability of the Intelligent Video Analysis (IVA) system. We carried out comprehensive experiments to show the proposed method on two public and one private data sets and demonstrate that the suggest approach beats various widely used baselines by a significant margin. This superiority enables to deploy this architecture in real-world commercial platform that proof-of-concept (POC) experiments were carried on Microsoft Azure cloud to identify its significance.

The main contributions of the proposed approach are summarized as follows.

- For detecting a person in a video surveillance system, an end-to-end hybrid architecture between an intelligent camera and a cloud server was proposed. It is a framework that is iterative which captures useful data for ongoing comprehension and learning.
- For the intelligent front-end, we suggest a Multi-Scale ResBlock (MSRB)-PeleeNet for our object detection model in the intelligent front-end camera motivated and use the ResBlock of the residual structure at each feature map.
- We are creating a novel domain adaption strategy in the inference phase, resulting in a unique personal model, to accurately comprehend each space and contextual data in intelligent back-end system.
- We conducted POC testing in real-world settings and lab studies on the set of test data to evaluate and verify the suggested technique in a more usable manner.

II. ARCHITECTURE DESIGN OVERVIEW

We propose a novel architecture for video surveillance which combines two different intelligence systems as depicted in Fig. 2. Unlike traditional server-based video surveillance structure, we deploy a detection model on the camera as an intelligent front-end architecture to permit the scalability, adaptability, and expansion of the system.

To create a reliable initial generic object detection model on the edge camera, we train the CNN model with training set from several sources of video surveillance. A thorough evaluation of the data set having been completed, initial model is deployed on cameras for detecting an important object, as shown in Fig. 2. Subsequently, learning the precise spatial and context information, the original model is replaced to an individual-specific model. If an image's confidence score is high according to the present model, we use it as a verification data set when developing another IVA functions by defining dissimilarity metric (DSM), as illustrated in Fig. 2. On the contrary, when it shows a small confidence value below than pre-defined threshold, it is utilized as a potential option for a new training data set to get around the uncertainness (i.e., to update the model).

It is crucial that the present model utilize new data as part of an ongoing adaption process via model version manager, as shown in Fig. 2. Additionally, our architecture iteratively carries out the model assessment method to solve the catastrophic forgetting issue and enable ongoing learning in real-time. It strengthens the adaptability of our system to new, specific target data.

III. NOVEL FRONT- AND BACK-END INTELLIGENCE FOR LIFELONG LEARNING

3.1. Intelligent Front-End Architecture

It is crucial to develop a model not largely dependent on a particular layer to deploy a CNN model to a camera (e.g.,

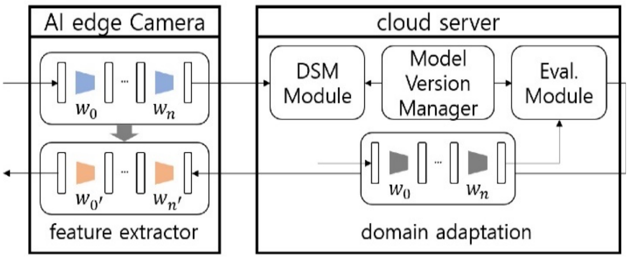

Fig. 2. Proposed architecture for intelligent video surveillance.

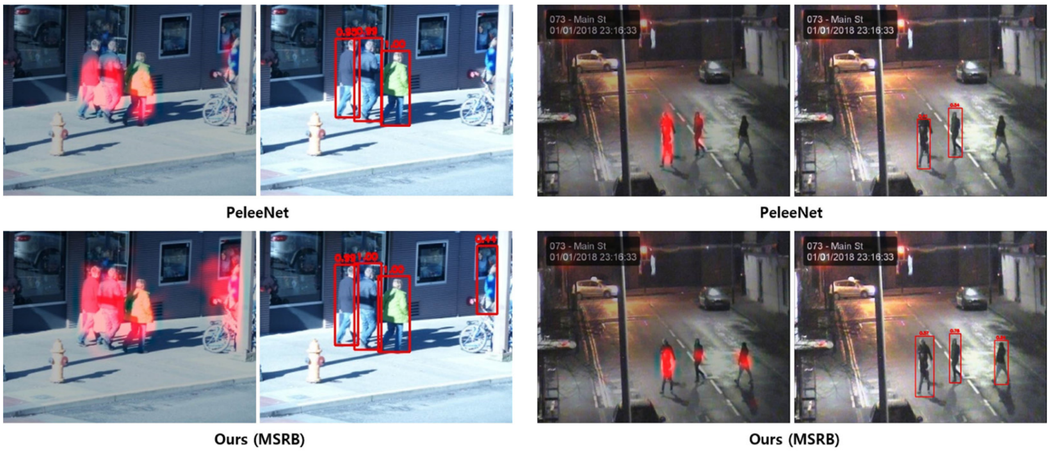

Fig. 3. The results of PeleeNet and our CNN model (MSRB) in the difficult cases of (left) the occluded person and (right) the blurred person.

depthwise separable convolution layer). The ResBlock used in the original model of PeleeNet applied only one (3×3) size kernel, so that it is difficult to robustly detect various sizes of people especially in occluded and blurred areas shown in Fig. 3. It is important to consider both the surrounding information and the object itself when determining whether objects are obscured by other objects or the image borders. The inability to extract and examine the pixel data surrounding the objects while employing a common kernel on the feature map. This problem is more severe when the resolution of image is lower, which is usually the case in cloud-based video surveillance services.

In general, some object detectors like Yolo v5 [10] and ssFPN [11] were proposed to detect objects more accurately. However, a novel object detection model which efficiently captures important objects in real-time has not yet been systematically developed especially for video surveillance camera. Thus, we suggest a Multi-Scale ResBlock (MSRB)-PeleeNet for our object detection model in the intelligent front-end camera motivated and use the ResBlock of the residual structure at each feature map, as shown in Fig. 4. The MSRB proposed is made up of different sized kernels (3×3, 5×5, and 7×7) to handle the surrounding context information more concretely. Multi-scale kernels can be used to evaluate each of the layer characteristics, yielding statistically better performance at various scenes. It is also noticeable that the performance is maintained by using the retrieved context information even in severe blurring cases.

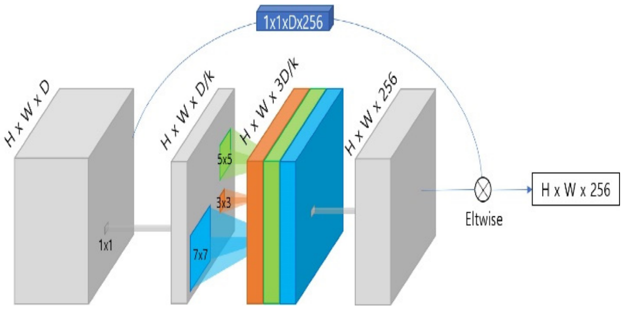

Fig. 4. Network architecture of object detection model.

Fig. 3 shows the qualitative results on two video surveillance scenes, where the proposed object detector detects a person more accurately even in the occluded or blurred cases compared to the original PeleeNet. Furthermore, we disregard the last three feature maps (5×5, 3×3, and 1×1) in the original design to lower the expensive computational expenditures in PeleeNet. Finally, we can identify that MSRB-PeleeNet is appropriate to detect objects for front-end surveillance camera due of its high accuracy and simplicity of computation.

3.2. Intelligent Back-End System

For intelligent back-end architecture, new domain adaption scheme is what we suggest on MSRB-PeleeNet at each camera to become familiar with the appropriate personal surroundings. Denote data of the target during the adaptation in frame duration N as $x_A^t = \{x_0^t, x_1^t, \ldots, x_N^t\}$, where n is the frame number within $0 \leq n \leq N$. Even if there isn't any labeled data that corresponds to the bounding box coordinates for training, we may use it as a negative set when there is no one present on the image.

Here, we provide a novel technique for sampling the negative background images automatically during the adaptation period. First, we choose the first frame x_0^t as an element of the final selected set x_S^t $(x_S^t \subset x_A^t)$. In this step, even there are some foreground objects in the scene, even after many rounds, we can still provide effective outcomes for domain adaptation. The following step is to acquire the gray-scale frame difference image, $\hat{x}_{D_n}^t = \hat{x}_n^t - \hat{x}_{n-1}^t$, where $\hat{x}_n^t$ is the gray-scale image in the n^{th} frame. Hence, if the average of difference across the most recent few frames (5 frames in this paper) is more than a specific threshold ε, it means moving objects appear on the n^{th} frame. Avoiding the repetition problem is what we choose a $(n-1)^{th}$ frame, x_{n-1}^t as an alternative for the consecutive step. It is preferable to choose only one frame as a representative image instead of choosing from all the candidates in a surveillance video since the next frames are nearly similar.

The third phase is where we define a new metric (*M*) of x_{n-1}^{t} through all frames in x_{S}^{t} to avoid the repetition in this set by considering the luminance, contrast, structure, and texture dissimilarity:

$$\boldsymbol{M(x_{n-1}^{t}) = ssim(x_{n-1}^{t}) \times \frac{1}{1+\exp(-\daleth(x_{n-1}^{t}))}}, \tag{1}$$

where $ssim(\cdot)$ is the structural similarity index metric (SSIM) [12], which is a remarkable human visual system-aware objective image quality assessment method that it is calculated with all images in x_{S}^{t}, and $\daleth(\hat{x}_{n-1}^{t})$ is the absolute difference of texture entropy on $(n-1)^{th}$ frame, $E_{(\hat{x}_{n-1}^{t})}$, which means the dissimilarity of visual entropy of texture information based on the information theory [13].

We only use the negative set x_{S}^{t} when refining the first object detector for training without any positive samples for detection. As a result, if the initial detector discovers bounding boxes that are considered as false positives, since this area is devoid of any thing in the frame, according to the confidence score of, loss value is penalized on the predicted bounding box as follows.

$$L(c,l) = \frac{1}{N_p}\{L_{conf}(c) + L_{loc}(c)\}, \tag{2}$$

$$L_{conf}(c) = -\sum_{i}^{N_p} \log(c_i), \tag{3}$$

$$L_{loc}(c) = \sum_{i}^{N_p} smooth_{L1}(l_i), \tag{4}$$

where N_p is the quantity of expected bounding box l and c is the confidence score.

IV. EXPERIMENTS

We carried out two different sorts of studies to confirm the correctness and applicability of the suggested strategy: one in the laboratory (Section 4.1), and the other through a POC experiments (Section 4.2).

Please be aware that in the studies, we employed every object detector model available to identify people in situations of video surveillance. Hence, all the data sets included in this study are video clips that were shot in fixed places and had small fields of view.

4.1. Experimental Set-up

In our research, we took use of two publicly accessible data sets of Performance Evaluation in Tracking for Surveillance (PETS) 2009 [14] and Oxford Town Centre [15]. In the PETS 2009 data set, we used the 'S2-L1' video of a scenario outside with several people crossing each other, totaling 795 frames. Furthermore, we conducted the experiments on the Oxford Town Centre data set displaying people strolling through a busy street having 25 fps with a Full HD resolution. However, there is a limited amount of real-world video surveillance data available to use due to the privacy regulations. Thus, we utilized a private data set made up of abount 130*K* training images (Private_T) and 6*K* evaluation images (Private_E) as shown in Fig. 5.

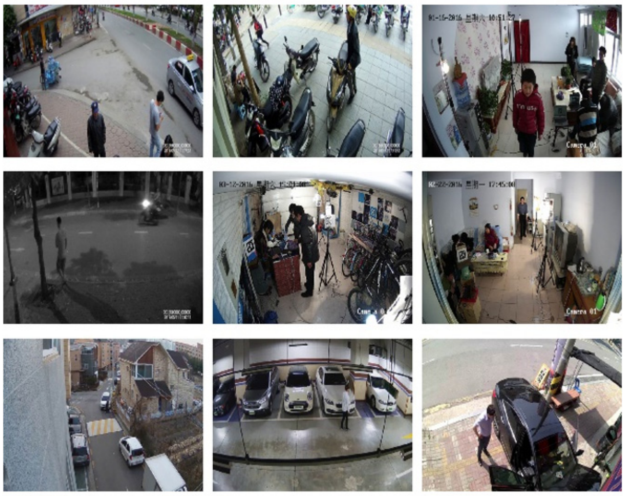

Fig. 5. Examples of private dataset.

In the training and domain adaptation of the object detector, the batch size and learning rate were specified as 64 and 0.001, respectively. When trained with the private data set of Private_T, the network was trained over $210K$ iterations having the learning rate of 10^{-3}, the weight decay of 5×10^{-4}, and the learning rate decay value of 0.1 at $90k$ and $120K$. A particular size was applied to the inputs for those benchmark techniques, and training was done with the default options utilized in each algorithm's official, publicly available code. We adopted SSD [16] as a baseline backbone in the case of MobileNet v3 [17] and PeleeNet [18]. K-means clustering was used to obtain the anchor size for each training data sets.

The tests were carried out using a Qualcomm Visual Intelligence Platform with QCS603, which is intended for effective machine learning on Internet of Things (IoT) devices. Mean Average Precision (mAP), extensively utilized in earlier research on object detection, is employed with an Intersection over Union (IoU) threshold of 0.5 to compare accuracy.

4.2. Experimental Results and Discussion

To show the performance of MSRB-PeleeNet, we evaluated its effectiveness with several earlier object detection models of YOLO v4 [19], MobileNet v3, PeleeNet. We measured mAP values with the processing time on the camera

Table 1. Results of proposed method.

Method	DB	Image size	fps	Mean average precision		
				PETS2009	Oxford	Private_E
Yolo v4 (+D.A.)	Private_T	416×416	2.5	0.8341 (+0.0525)	0.7214 (−0.0048)	0.7952 (+0.1002)
MobileNet v3 (+D.A.)		416×416	48.8	0.3121 (+0.3892)	0.1884 (+0.4983)	0.3097 (+0.4241)
PeleeNet (+D.A.)		512×384	30.9	0.7886 (−0.0215)	0.6036 (+0.1187)	0.7132 (−0.0668)
Ours (MSRB) (+D.A.)		512×384	28.8	0.8105 (+0.0892)	0.7392 (+0.1501)	0.8603 (+0.0552)

after training with same data sets for detecting a people in video surveillance sequences.

Table 1 summarizes the performance of mAP and frame per seconds (fps) on two publics and one private data set on Digital Signal Processor chip. Our suggested model exhibits the greatest mAP values and a suitable processing time by a significant margin on an edge camera. Compared to the original PeleeNet, our model performs more accurately across all test data sets but at a slightly more computation time about 2−3 ms. In addition, we would like to underline that, when compared to the outcomes of the original object detector, our suggested domain adaption method's efficacy is noteworthy. The fps value is decreased in all models due to the short adaptation time required for understanding context information, however it is insignificant for inference in the actual world. The updating of the original model requires the most work, which is caused by fine-tuning.

4.3. Proof-of-the-Concept Tests

We conducted POC testing on a commercial video surveillance service, which manages about 130*K* video cameras, to validate if the suggested technique functions in a real-world context. Three hundred video cameras were used (for a week) among those to validate our method.

All the implementations of the proposed back-end system were deployed on Microsoft's Azure services. Every camera has a container-based client to interact with the server that records an image and its associated meta data (object detection results), together with their identity data, were transmitted to the storage server.

Fig. 6 plots the exemplar result frames for four representative test sequences that were validated from the test sequences. The results show several false positives alongside a person or at the backgrounds, as indicated by the first column in Fig. 6. In addition, false positives were more frequent than true negatives, and those errors were more likely to occur in the infrared mode of the camera. The following column in Fig. 6 displays the outcomes of domain adaptation, where the accuracy of person detection was maintained but the number of false positives was significantly decreased.

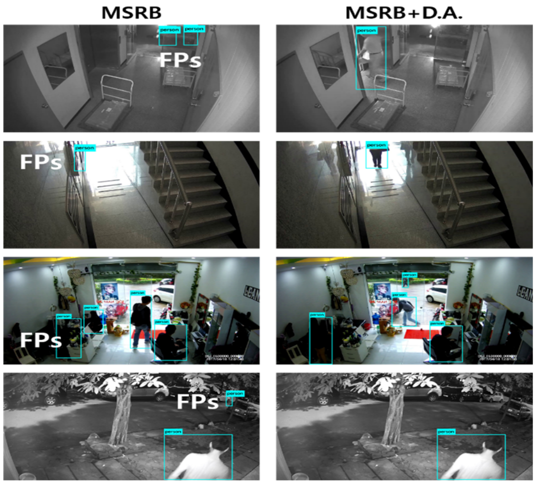

Fig. 6. POC results on four different examples.

V. CONCLUSION

We suggest an innovative end-to-end system architecture for continuous and gradual accuracy gain of video analytic algorithms, consisting of a hybrid combination of front-end and back-end intelligence. It is a continual process where the current model is self-evolving using context data in a continuous process of adaptation while maintaining the system's scalability, adaptability, and expandability. We anticipate that the suggested strategy will be crucial in enhancing commercial platforms with lifelong learning as intelligent video analytics continue to grow.

REFERENCES

[1] B. Bogdoll, M. Nitsche, and J. M. Zöllner, "Anomaly detection in autonomous driving: A survey", in *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Jun. 2022, pp. 4488-4499.

[2] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy,

"A review on deep learning in medical image analysis", *International Journal of Multimedia Information Retriecal*, vol. 11, no. 1, pp. 19-38, Mar. 2022.

[3] B. Kwon and T. Kim, "Toward an online continual learning architecture for intrusion detection of video surveillance," *IEEE Access*, vol. 10, pp. 89732-89744, 2022.

[4] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection", in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[5] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[6] S. H. Kim, J. H. Choi, T. K. Kim, and C. I. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection", in *IEEE International Conference on Computer Vision* (ICCV), Oct. 2019.

[7] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 2, pp. 361-374, Feb. 2014.

[8] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 6092-6101.

[9] K. Karoudis and G. D. Magoulas, "An architecture for smart lifelong learning design", in *Innovations in Smart Learning,* Sinapore, pp. 113-118, 2017.

[10] Y. Ma and P. Ren, "Lightweight object detection algorithm based on YOLOv5 for unmanned surface vehicle," *Frontiers in Marine Science,* vol. 9, p. 2585, 2023.

[11] H. J. Park, Y. J. Choi, Y. W. Lee, and Kim, B. G. "ssFPN: Scale sequence (S^2) feature based feature pyramid network for object detection," *arXiv preprintarXiv: 2208.11533*, 2022.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity"*, IEEE Transactions on Image Processing,* vol. 13, no. 4,pp. 600-612, Apr. 2004

[13] T. Kim, J. Kim, S. Kim, S. Cho, and S. Lee, "Perceptual crosstalk prediction on autostereoscopic 3d display", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27 ,no. 7 ,pp. 1450-1463, Jul. 2017.

[14] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge", in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Dec. 2009, pp. 1-6.

[15] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video", *CVPR 2011*, pp. 3457-3464, Jun. 2011.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C. Y. Fu, et al., "SSD: Single shot MultiBox detector", in *Proceeding European Conference Compututer Vision (ECCV),* Oct. 2016, pp. 21-37.

[17] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, and M. Tan, et al., "Searching for MobileNet V3", in *Proceeding IEEE/CVF International Conference on Computer Vision (ICCV),* Oct. 2019, pp. 1314-1324.

[18] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices", in *Proceeding Conference Neural Information Processing Systems (NeurIPS),* Dec. 2018, pp. 1-10.

[19] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection", *arXiv:2004.10934*, 2020.

AUTHOR

Taewan Kim received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively. From 2015 to 2021, he was with the Vision AI Laboratory, SK Telecom, Seoul. In 2022, he joined as a faculty with the Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, where he is currently an Assistant Professor. His research interests include computer vision and machine learning including continual and online learning.