IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2021, accepted July 16, 2021, date of publication July 20, 2021, date of current version July 27, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3098713*

# Domain Adaptation Deep Attention Network for Automatic Logo Detection and Recognition in Google Street View

**ERVIN YOHANNES[1], CHIH-YANG LIN[2], (Senior Member, IEEE), TIMOTHY K. SHIH[1], (Senior Member, IEEE), CHEN-YA HONG[1], AVIRMED ENKHBAT[1], AND FITRI UTAMININGRUM[3]**

[1]Department of Computer Science and Information Engineering, National Central University, Taoyuan City 32001, Taiwan
[2]Department of Electrical Engineering, Yuan-Ze University, Taoyuan City 32003, Taiwan
[3]Faculty of Computer Science, University of Brawijaya, Malang 65145, Indonesia

Corresponding author: Chih-Yang Lin (andrewlin@saturn.yzu.edu.tw)

**ABSTRACT** Signboards are important location landmarks that provide services to a local community. Non-disabled people can easily understand the meaning of a signboard based on its special shape; however, visually impaired people who need an assistive system to guide them to destinations or to help them understand their surroundings cannot. Currently, designing accurate assistive systems remain a challenge. Computer vision struggles to recognize signboards due to the diverse designs that combine text and images. Moreover, there is a lack of datasets to train the best model and reach good results. In this paper, we propose a novel framework that can automatically detect and recognize signboard logos. In addition, we utilize Google Street View to collect signboard images from Taiwan's streets. The proposed framework consists of a domain adaptation that not only reduces the loss function between source-target datasets, but also represents important source features adopted by the target dataset. In our model, we add nonlocal blocks and attention mechanisms called deep attention networks to achieve the best final result. We perform extensive experiments on both our dataset and public datasets to demonstrate the superior performance and effectiveness of our proposed method. The experimental results show that our proposed method outperforms state-of-the-art methods across all evaluation metrics.

**INDEX TERMS** Signboard, detection, recognition, logo, domain adaptation, deep learning, channel attention.

## I. INTRODUCTION

In 1993, the concept of logo recognition was introduced into computer vision. However, object detection and recognition have remained a challenging issue within logo recognition since then [1] due to a lack of quantifiable definitions for logos. A logo can consist of text, images or a combination of both. It is the representation of a product or brand [2], which can be found on social media [3], television or print media, street views, etc. [4]. Logos often appear prominently on street signage, and serve as important cues for location navigation, especially on street view images. By looking at signage, residents and visitors can determine what services a shop provides without entering it. However, these signs are not helpful to the visually impaired, who cannot see or have never previously seen the signage. A solution to this issue is to provide the visually impaired with a robotic assistant that could help them check for obstacles on the road and their surroundings [5]–[7]. However, such a system must be able to properly detect the signage, and do so quickly and accurately to be useful.

Significant advances in deep learning have enabled the development of well-known frameworks based on state-of-the-art convolutional neural networks (CNNs) such as R-CNN, Faster R-CNN, SSD [8], YOLO (You Only Look Once) [9], and DenseBox, which allow detection and recognition to be carried out in real time [10]. YOLOv2 has a

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

powerful detection network and has been successfully combined with other networks [7].

Recently, a Faster R-CNN proposed by [11] for automatic signboard detection in Bangladesh used low-quality images taken from Google Street View. This Faster R-CNN consisted of a base CNN, an RPN, non-maximum suppression, ROI Pooling, and an R-CNN. It was used for image preprocessing and hyperparameter tuning. The highest mean average precision (mAP) score reached 82% and it was categorized into two classes: signboard and non-signboard [12]. In [13], YOLOv2 was implemented into automatic signboard detection in Taiwan street views. The method not only detected signboards, but also collected datasets from the street view. The datasets contained signboard images of a fixed size from several convenience stores in Taiwan.

Signboards exist everywhere to attract people's attention or to call attention to something important. However, this information cannot be read by the visually impaired. Although deep learning has been successfully used in many fields [8], [9], [12]–[14] and shows promise in assistive technologies for the visually impaired, its success heavily depends on the sufficiency of datasets. As types of signboards vary from country to country, the public signboard dataset is quite limited and cannot be used in different countries. We propose an automatic signboard detection and recognition system in this paper based on the domain adaptation deep attention network to integrate existing signboard datasets and adapt to another new signboard domain. To address the lack of signboard datasets in Taiwan, a tool was developed based on Google Street View to collect possible signboards around Taiwan's streets. The overall idea is outlined in Figure 1.

The main contributions of this paper are that:

1) We propose a novel framework for automatic signboard detection and recognition in similar feature domain representation using domain adaptation method as our pretrained model combined with non-local blocks and attention mechanism.

2) We provide the largest public signboard dataset[1] with 29,727 images in 14 classes. The current largest signboard dataset is BSVSO with 2,043 images, but BSVSO only provides 1 class (with or without signboards).

3) Based on the experimental results on available datasets, our proposed framework has superior performance compared to state-of-the-art methods.

The remainder of this paper is organized as follows: In Section II, several related works from existing literature are described. In Section III, a detailed design of our system is presented. Experimental results and a discussion of our system are provided in Section IV. Finally, concluding remarks are given in Section V.

## II. RELATED WORK

Many attempts have been made to incorporate attention mechanism into CNNs and have been shown to improve detection and recognition systems. Moreover, domain adaptation has advantages in feature representation not only for similar domains, but also in limited class datasets. In this section, we discuss the detection methods, domain adaptations, and attention networks related to our proposed method.

### A. DETECTION METHOD

Object detection in deep learning can be roughly divided into two approaches. One is two-stage detection, such as in region-based convolutional neural network. The other is one-stage detection, such as Single Shot MultiBox Detector [8] and the YOLO series [9]. Two-stage detection first chooses region proposals and then classifies each proposal. One-stage detection only needs one network to simultaneously predict bounding boxes and classification [15]. Two-stage detection has higher accuracy, while one-stage detection is faster.

Single Shot MultiBox Detector [8] is a famous one-stage detection method. The main idea in this method is to utilize multiple classifiers from multiple scales of feature maps. This concept is motivated by feature pyramid networks [16]. The basic feature extracting network is VGG-16, which adopts anchor boxes in different feature maps of different layers. Each layer utilizes a different number of anchors, and the anchor size is set in advance. In the training phase, the method introduces hard negative mining to speed up convergence because there are usually many more negative samples than positive samples in nature. In addition, the method performs data augmentation to prevent overfitting. Because the architecture is an end-to-end network, it has high speed both in training and testing while maintaining great accuracy.

Although Faster R-CNN achieves good accuracy, it is not an end-to-end network and needs multistage training. Therefore, Redmon *et al.* [9] proposed YOLO, another famous one-stage detection method. YOLO uses a single end-to-end network to predict both bounding boxes and category probabilities with only an input image. Redmon *et al.* viewed detection as a regression problem, so they measured the loss by mean square errors. YOLO includes several convolutional layers and fully connected layers. An input image is divided into 7 × 7 grids, and each grid is responsible for predicting one kind of object. Each grid predicts a fixed number of bounding boxes. The final output is the probability of each category for each grid, the confidence score, and the information for each bounding box, including the coordinates of the center point and its height and width. The end-to-end network structure allows YOLO to be very efficient.

Compared to Fast R-CNN, YOLO suffers from more localization errors because it has fewer bounding boxes. To improve accuracy and speed, Redmon and Farhadi [13] suggested a variety of concepts, such as batch normalization after convolution, multiscale training, and a convolutional network with anchor boxes, to improve upon YOLO in YOLOv2. Multiscale training was applied to adapt to different input dimensions. The concept of anchor boxes was borrowed from Faster R-CNN, while the size of anchors

[1] Available at "https://github.com/ervinyo/Signboard-datasets".

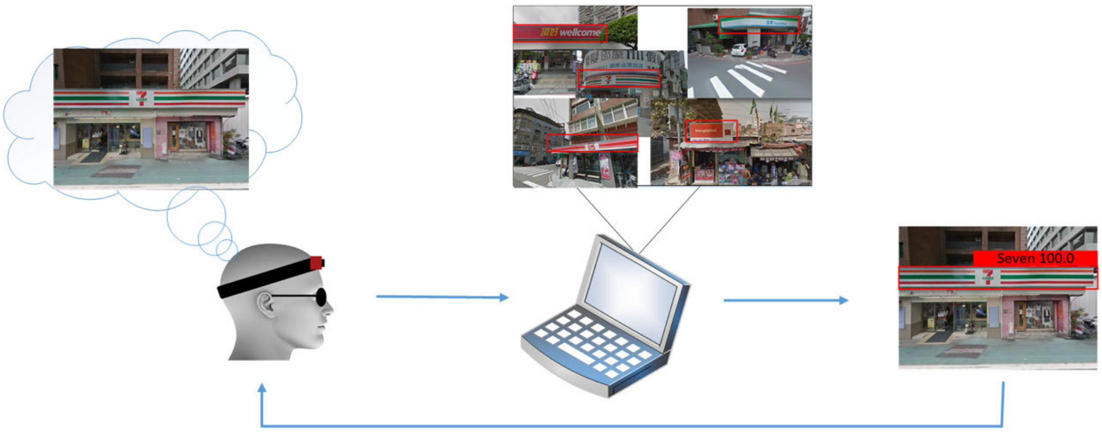

**FIGURE 1.** The guiding system. The visually impaired user wears a camera on his/her head and tells the system a target destination. The camera receives images and gets the recognition result to help the visually impaired user walk toward the destination.

was obtained via K-means clustering. The researchers also added a pass-through layer to pass previous features to the last feature map to predict both large and small objects well. To speed up training and testing, they discarded VGG-16 and designed a new classification model called Darknet-19 to avoid floating-point calculations. They additionally employed global average pooling to replace the fully connected layer.

YOLOv3 [14] was introduced to further increase accuracy in small object detection. The team borrowed the concept of the Single Shot MultiBox Detector [8] to predict objects on multiple scales of feature maps and proposed a new network architecture called Darknet-53, which was similar to Darknet-19 but used a simple residual block to replace the original convolutional layers. YOLOv4 [14] added attention networks to the loss function, but did not obtain better accuracy. In our experimental results, the basic Darknet-53 did not achieve higher accuracy than the basic Darknet-19.

### B. DOMAIN ADAPTATION

The domain adaptation technique aims to reduce the unimportant features in a domain shift, and has already been incorporated into deep learning and mapping both domains into a feature space [18]. Moreover, domain adaptation is a well-known and useful research field in transfer learning because it can be augmented by generative adversarial networks (GANs) [19], which are designed to narrow the gap between cross-domain and feature level adaptation so that the results are enhanced. He *et al.* [20] reported on domain adaptation for classifier-level adaptation, and their results increased not only the domain-invariant features, but also the auxiliary information on class. In addition, Teng *et al.* [21] designed a classifier with a deep adversarial domain adaptation for classifying remote sensing images. They adapted the concept of deep learning to reach the best classification results while preventing misalignment [22], [23].

Domain adaptation can also be implemented in pedestrian and object detection systems. Pedestrian detection exerts maximum independence domain adaptation combined with transfer component analysis to receive the source and targeting domain data in a new space, but the distributions of the two domains are almost the same. Moreover, domain adaptation could adapt domain shifts in appearance and semantic levels of an object to achieve higher accuracy and effectiveness [24], [25]. Oliveira *et al.* [26] proposed a new method of domain adaptation named the Conditional Domain Adaptation Generative Adversarial Network (CoDAGAN) for segmentation purposes in biomedical images. It merges unsupervised [27] and supervised networks to become a semi-supervised method that can learn from unlabeled and labeled data. The proposed method achieved better results than other state-the-art methods in labeled data scenarios. Xiao *et al.* studied domain adaptation for enhancing depth images, which could enable domain transfer between a simulator and a real camera since it generates two outputs: a degradation part and an enhancement part [28].

Li *et al.* [29] has proposed a domain adaptation named Adversarial Tight Match (ATM) that uses adversarial training and metric learning to reduce the loss of the divergence between the two domains (e.g. source and target). His approach differs from traditional requirements by minimizing the inter-domain divergence and maximizing intra-class density. This method is able to achieve state-of-the-art performance by joining two levels: feature and sample levels, to improve the knowledge during the training phase on the labels with low confidence scores [30]. It can minimize marginal and conditional mean discrepancy (MMD) and re-weight instances by landmarks selection. This technique can be implemented in both homogeneous and heterogeneous domains. Gao *et al.* [31], [32] have proposed cross-domain techniques that can prove the domain adaptation issues by using pairwise attentive adversarial spatiotemporal network and pairwise two-stream convolution network.

The benefits include established higher domain discrepancies and improved discrimination on spatiotemporal features in order to effectively learn domain invariant features.

### C. ATTENTION NETWORK

Attention mechanisms have been adapted to different domains because of their ability to model dependencies. The main concept of an attention mechanism focuses on important features to match the input data. Recently, many attention network models have been applied in various fields. Zheng *et al.* [33] proposed a graphic multiple attention network and applied an encoder-decoder architecture to predict traffic lights. Moreover, attention networks can boost the results of a CNN model with a convolutional block attention module comprised of a channel attention module and a spatial attention module [34]. An attention network can also be applied to segmentation, detection, and recognition tasks [35], [36]. Segmentation can be enhanced in terms of feature extraction by using two attention networks, including semantic interdependencies and channel dimensions, which are the results of more precise segmentation. Meanwhile, a spatial-channel combinational attention mechanism was introduced in pavement crack detection with the objective of improving important feature representation. For recognition, Canjie *et al.* [37] presented a multi-object rectified attention network that achieved state-of-the-art performance in scene texts. The system combined a multi-object rectification network and an attention-based sequence network.

In recent years, attention networks have been able to integrate attribute features and body part classification and could be applied to personal re-identification, such as personal ID tasks, detection tasks, and crucial detection tasks [38]. Attention networks can improve the super-resolution of an image, which is also useful for checking deeper and wider layers, since super-resolution has its flaws. An attention network has more power to extract layers and the important information contained inside the layers [39]. Zhang *et al.* [40] reported on dual attention networks for object counting. This method was effective on pyramid structures and provided the spatial attention for processing multiscale features on a large scale, thus improving object counting performance.

## III. METHODS

We used source and target datasets to gain the best pre-trained model based on the domain adaptation. In the domain adaptation, the source and target domains have a similar feature space. Here, a source is a set of labeled data that contains more than one class. A target is a set of data (usually unlabeled) with a different distribution than the source, but contains similar types as the source. In our proposed method, we use both our dataset (source) and BSVSO (target) for domain adaptation. There are 1,400 source images and 566 target images.

Domain adaptation computes the shift in distribution through domains so that relevant knowledge can be transferred from source to target in order to predict target labels. It has the advantage of reducing the loss between the source and the target since it already obtains domain features for both sides.

**TABLE 1.** DarkNet-53 network.

| | Type | Filters | Size | Output |
|---|---|---|---|---|
| | Convolutional | 32 | 3 x 3 | 256 x 256 |
| | Convolutional | 64 | 3 x 3 / 2 | 128 x 128 |
| 1x | Convolutional | 32 | 1 x 1 | |
| | Convolutional | 64 | 3 x 3 | |
| | Residual | | | 128 x 128 |
| | Convolutional | 128 | 3 x 3 / 2 | 64 x 64 |
| 2x | Convolutional | 64 | 1 x 1 | |
| | Convolutional | 128 | 3 x 3 | |
| | Residual | | | 64 x 64 |
| | Convolutional | 256 | 3 x 3 / 2 | 32 x 32 |
| 8x | Convolutional | 128 | 1 x 1 | |
| | Convolutional | 256 | 3 x 3 | |
| | Residual | | | 32 x 32 |
| | Convolutional | 512 | 3 x 3 / 2 | 16 x 16 |
| 8x | Convolutional | 256 | 1 x 1 | |
| | Convolutional | 512 | 3 x 3 | |
| | Residual | | | 16 x 16 |
| | Convolutional | 1024 | 3 x 3 / 2 | 8 x 8 |
| 4x | Convolutional | 512 | 1 x 1 | |
| | Convolutional | 1024 | 3 x 3 | |
| | Residual | | | 8 x 8 |
| | Avgpool | | Global | |
| | Connected | | 1000 | |
| | Softmax | | | |

Next, we employed the pre-trained model in our detection and recognition networks for detecting and recognizing the object. The detection and recognition network is the primary network since it combines certain models as one. Additionally, the network consists of DarkNet-53 for the backbone as shown in Table 1, nonlocal blocks and channel attention networks. DarkNet is a popular backbone due to its well performance and simple network architecture, and it is easy to change the architecture to fit our needs. Based on DarkNet, we utilized residual blocks, skipping layers, additional layers, and detection layers. Our residual blocks have 13 layers, which extract image features. There are 2 skipping layers, and we put the skipping layers on the 36th and 61st layers. The details of our proposed method can be seen in Figure 2.

First, we learn a mapping to an ordinary feature space, so that we can insert a domain-invariant vector into our discriminator. We use a discriminative model to decide whether the images belong to the source domain or the target domain and try to learn the mappings from both domains. The loss function for the discriminator is dependent on the target distribution. Basically, optimizing the loss function tries to label target distribution images as if they belong to the source domain. For the mappings, untied weights are used separately for the source and target domains since the features learned in one need not be same in the other due to the different domains. Two different CNNs are used to learn the mappings. We first train the source CNN and source classifier using

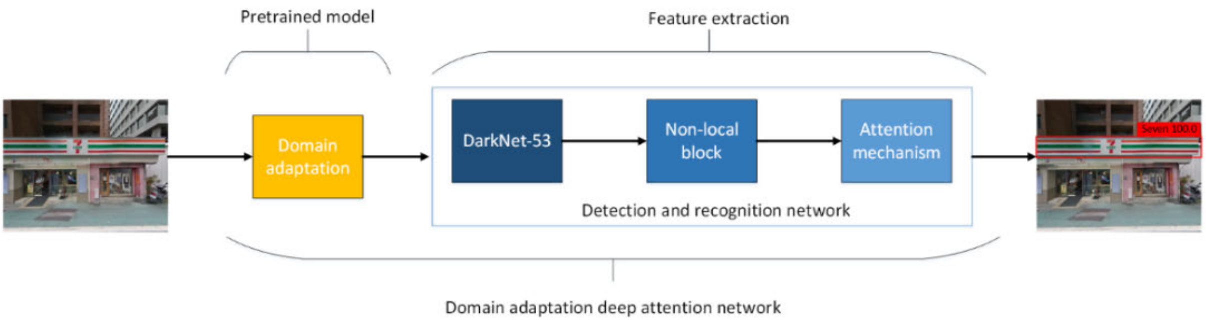

**FIGURE 2.** Flowchart of our proposed domain adaptation deep attention network.

typical image classification techniques. Then, we fix the source CNN and classifier after this step. In the domain adaptation step, we train a discriminator and the target CNN through the common adversarial process. The source mapping and the target mapping are learned in this stage.

We thus have two proposed methods: Proposed method I uses a nonlocal block and attention mechanism called deep attention as an addition feature extraction in our detection and recognition system. Proposed method II uses domain adaptation, non-local blocks, and attention mechanism to take advantage of the domain representation between the source and target datasets and combines it with deep attention results.

The attention mechanism can learn which position or channel is more important and which needs to attend to some features. With channel attention, some channels will assist in obtaining the result, which is implemented in different ways. The channel attention in our network architecture is introduced as a CBAM [34] as seen in Figure 3. First, it performs max pooling and average pooling on the input feature map separately. Next, both pools will pass the shared multilayer perceptron and initiate an elementwise sum, followed by an activation function. Finally, elementwise multiplication is initiated with the original input feature map. We add channel attention after the dilated block.

### A. PRETRAINED MODEL IN DOMAIN ADAPTATION

A dataset typically requires a common pre-trained model to train datasets such as Imagenet, MS COCO, and Pascal VOC. We realize that long computation times are needed to achieve convergence. Our datasets have similar domains with public datasets so that the computation needed to achieve convergence is faster than using the common pre-trained model. Target and source datasets taken from both our and BSVSO datasets are used to train them using the pre-trained model. Moreover, the domain adaptation can improve the detection and recognition results due to the similar domain features in the two datasets. Meanwhile, the common pre-trained model has different domains with our datasets and BSVSO datasets, so we must train each of them. The pre-trained modeling process in the domain adaptation is shown in Figure 4.

### B. NONLOCAL BLOCK

Normal convolutional operations only process local information. If we need a larger receptive field, we usually exert several consecutive convolutional layers, which is not very efficient. A nonlocal block is introduced from nonlocal neural networks [41], and the structure is shown in Figure 5. Nonlocal operations will consider all possible positions instead of only nearby positions such as convolutional operations. In the nonlocal block, the input feature map will be divided into three parts, and each part will halve the dimension by a 1 x 1 convolution. One of the first two parts will transfer and perform the matrix multiplication with the other part, then use softmax on the feature map. After that, the third part will also be multiplied with the previous feature map. Multiplication is followed by a 1 x 1 convolution to increase its dimension to match the original feature map size and sum it elementwise. The nonlocal block will maintain the same dimension for the output feature map as the input one. This block can be easily inserted into deep learning architectures. We introduce the nonlocal block as the last operation on the feature map, which we extracted from the 13th convolutional layer. This is another kind of operation used to raise the receptive field in our network architecture.

### C. ADDITION FEATURE EXTRACTION USING ATTENTION NETWORK

We add the attention network after the first detection layer and residual blocks since it can improve the next layer. The process for adding features using the attention network consists of a skipping layer, an addition layer with nonlocal blocks, a detection layer, and residual blocks. In the first detection layer, the scales for detection and recognition are few and limited, so the easy-to-learn model can take the additional feature extraction that is missing in the first detection layer and implement it in the next layer. A comparison of detection layers is shown in Figure 6.

In Figure 6, we can see the detection is unsure of the outcome, as it could be an incorrect result and will be recognized as such until the last detection layer. The function of the attention layer is to filter and give the exact result in the first layer of detection, just after nonlocal blocks enhance the

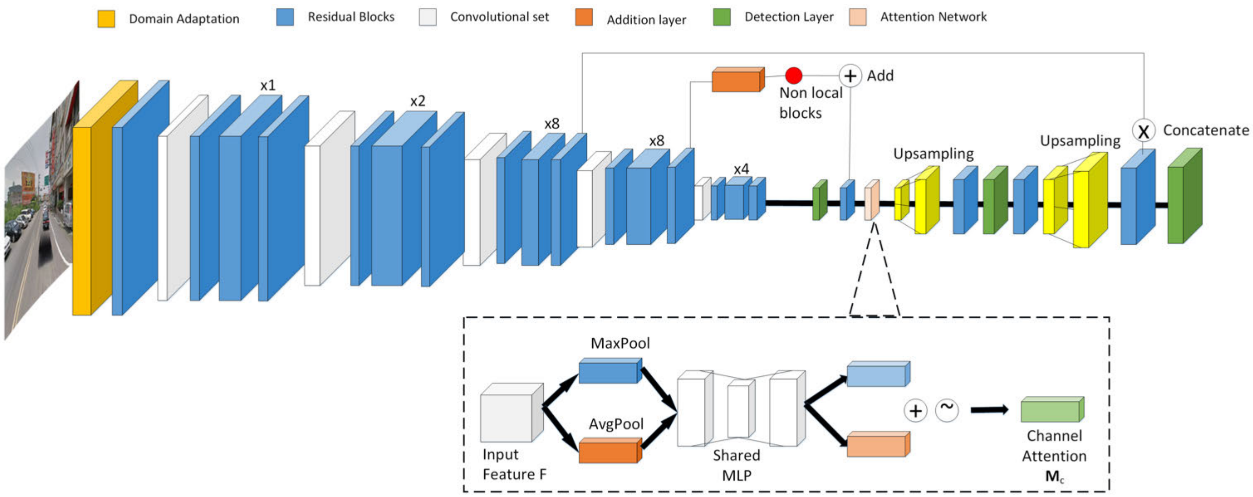

**FIGURE 3. Architecture of our proposed detection and recognition network.**

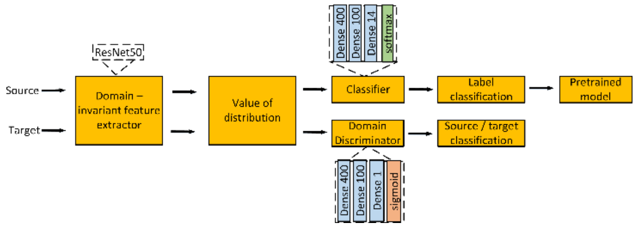

**FIGURE 4. Architecture of domain adaptation used in our proposed method. The input includes source and target datasets.**

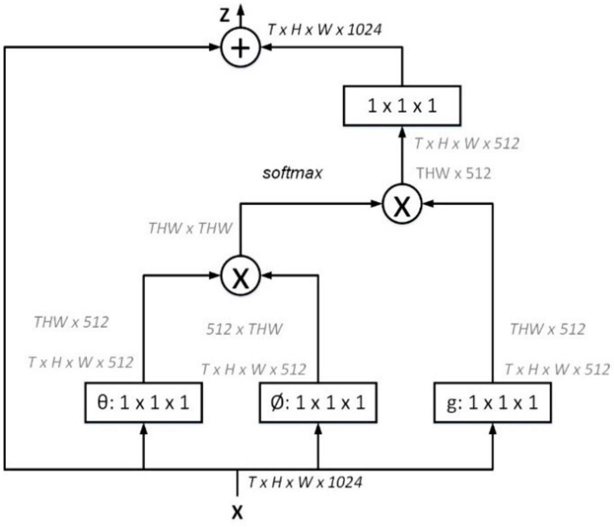

**FIGURE 5. Nonlocal blocks architecture.**

small objects and take the global information. The attention network can improve the detection and recognition results by choosing one result that is exactly the same as the domain feature weights. In Scale 1, the detection result consists of Hi-Life, Ok Mart, and McDonalds, which is the same as the result without the attention network. In Scale 2 with the attention network, the same signboard is detected as with the domain feature. However, without the attention network, the results are mixed. This means that the attention network is faster and obtains a more reliable result. Finally, Scale 3 shows the correctly detected sign; however, without the attention network, the results are mixed.

## IV. EXPERIMENTS

In this section, the performance of the proposed method on detecting and recognizing the object are evaluated. All experiments utilize the same testing environment and hyper-parameter settings.

### A. DATASETS

We used the BSVSO (Bangladesh Street View Signboard Objects) [10] dataset and our own dataset. The BSVSO dataset contains 2,043 images with VOC annotation format and a resolution of 1000 × 600 pixels. The dataset has one label class taken from 9 cities in Bangladesh, including Dhaka, Sylhet, Chittagong, Rajshahi, Khulna, Rangpur, Bogra, Pabna, and Barisan. The training and validation images are set to 1,429 and 614, respectively. Meanwhile, our dataset is a signboard dataset containing 29,727 images with VOC annotation format and a resolution of 500 × 400 pixels. Our dataset includes 14 classes of store logos, including Carrefour, Domino, Family Mart, Gas, Hi-Life, KFC, McDonalds, Mos Burger, Ok Mart, Post, Pxmart, 7-ELEVEN, Starbucks, and Wellcome, collected from 6 major cities in Taiwan. The training and validation images for this dataset are set to 23,786 and 5,941, respectively. Samples from both datasets are shown in Figure 7. We have created the largest dataset of signboards for detection and recognition issues. It is different from ReCTS-25k datasets [42] since ReCTS-25k only contains Chinese text on signboards and the dataset focuses on text recognition issues. On the contrary, our dataset is an image-based container and focuses on image detection and recognition issues.

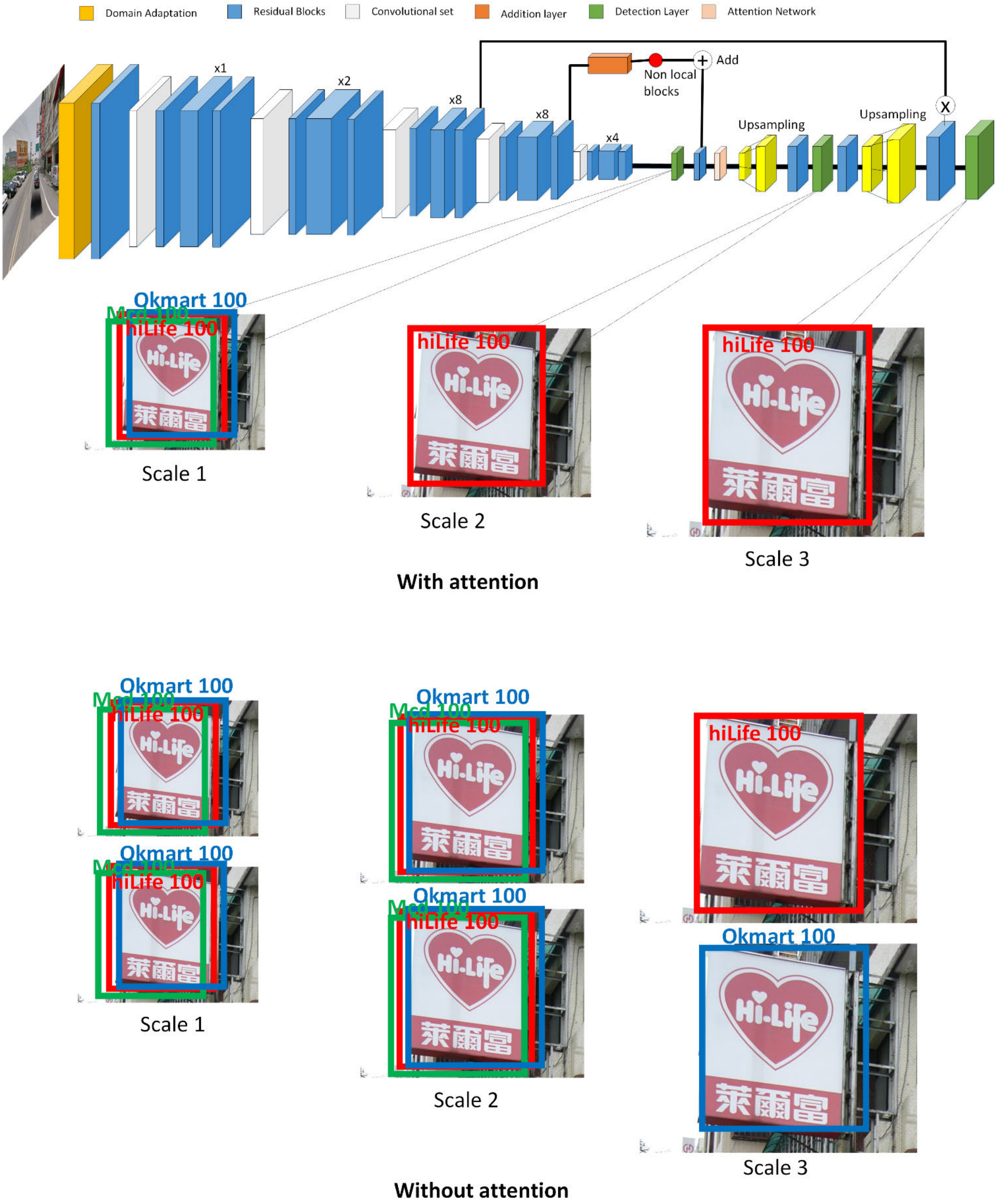

**FIGURE 6.** **Comparison of results with and without an attention network in the recognition and detection system.**

## B. EVALUATION METHOD

To evaluate our proposed method, we take the intersection over union (IoU) between two bounding boxes as the K-means distance for the purpose of obtaining a better IoU between the predicted bounding box and the ground truth. Directly taking the height and width of the bounding box as

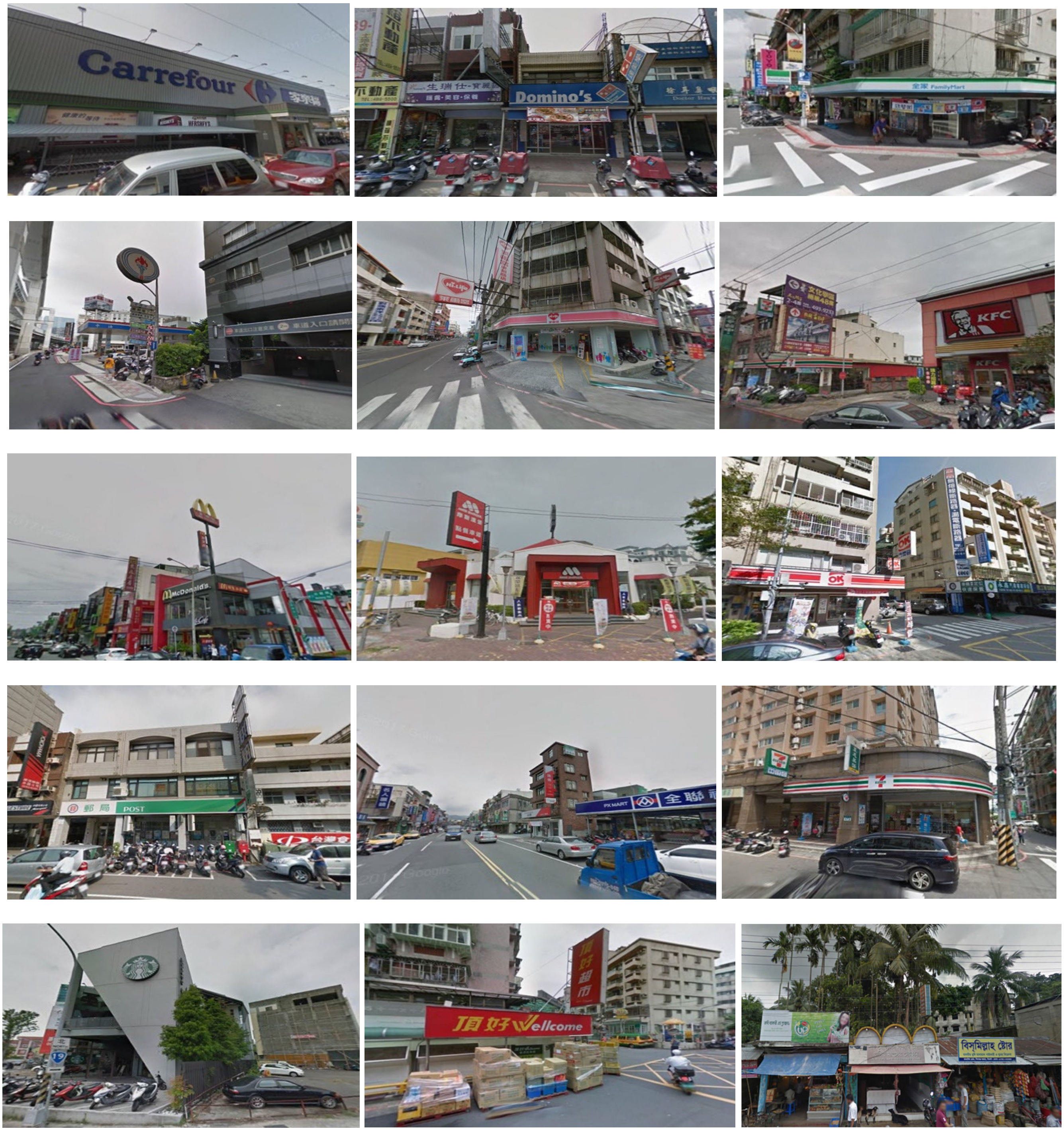

**FIGURE 7.** Dataset samples of our dataset in 14 classes (Carrefour, Domino, Family Mart, Gas, HiLife, KFC, McDonald, Mos Burger, Ok Mart, Post, Pxmart, 7-ELEVEN, Starbucks, and Wellcome) and Bangladesh Street View Signboard Objects (BSVSO) in one class (signboard).

the K-means distance would not be reasonable. The IoU is often used in estimating the accuracy in object detection and semantic segmentation. Because our dataset consists of signboards from Taiwan streets, the size of the bounding boxes is probably different from other common datasets. We should choose appropriate anchor sizes to train the network so that we can obtain better results.

### C. ACCURACY RESULTS BETWEEN SOURCE AND TARGET IN DOMAIN ADAPTATION

For domain adaptation, we tested both our dataset and the BSVSO datasets, as the domain adaptation needs images and has an annotation file. Here, we use 15 classes contains 14 classes of our dataset and one classes of the BSVSO dataset. We have 1400 images as source dataset and

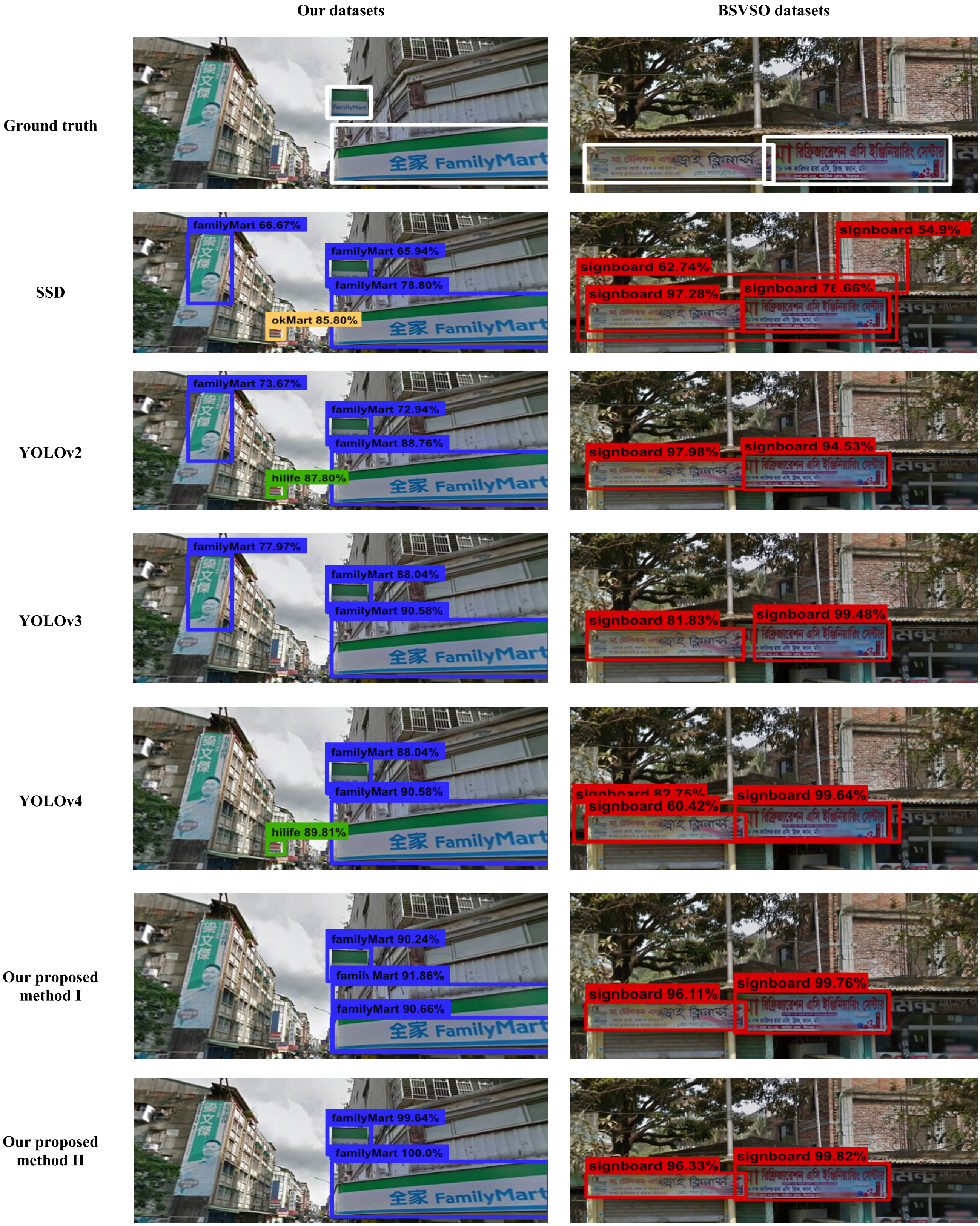

**FIGURE 8.** Results of logo detection and recognition using our datasets and BSVSO datasets. Our proposed method II shows the best results on both our and BSVSO datasets compared with other methods in terms of bounding box accuracy and confidence scores.

**TABLE 2.** Accuracy results between source and target in the domain adaptation method.

| Epoch | Source | Target |
|---|---|---|
| 500 | 71.72 | 41.28 |
| 1000 | 85.01 | 48.34 |
| 1500 | 89.39 | 52.06 |
| 2000 | 89.49 | 53.53 |
| 2500 | 91.06 | 56.57 |
| 3000 | 93.87 | 59.81 |
| 3500 | 93.99 | 59.97 |
| 4000 | **95.77** | **60.49** |
| 4500 | 95.60 | 59.08 |
| 5000 | 96.77 | 59.10 |

**TABLE 3.** Comparison of results from state-of-the-art detection & recognition methods and Proposed Method I on our dataset.

| Classes | SSD | YOLO v2 | YOLO v3 | YOLO v4 | Our proposed method I |
|---|---|---|---|---|---|
| Carrefour | 62.74 | 80.65 | 85.48 | 88.07 | **88.19** |
| Domino | 90.17 | 91.06 | 96.26 | 96.94 | **97.08** |
| Family Mart | 91.23 | 86.41 | 95.36 | 96.67 | **96.71** |
| Gas | 78 | 93.67 | 92.99 | 94.56 | **96.10** |
| Hi-Life | 76 | 86.59 | 95.01 | 94.73 | **95.34** |
| KFC | 87.93 | 85.81 | 92.27 | 91.89 | **92.16** |
| McDonald | 78.57 | 80.55 | 87.89 | 90.51 | **91.12** |
| Mos Burger | 66.67 | 82.14 | 85.70 | 89.54 | **90.46** |
| Ok Mart | 92 | 88.08 | 91.34 | 93.31 | **93.34** |
| Post | 76 | 90.15 | 91.23 | 91.43 | **91.83** |
| PxMart | 86 | 89.64 | 90.52 | 92.61 | **93.08** |
| 7-Eleven | 80.93 | 85.07 | 95.12 | 94.61 | **94.68** |
| Starbucks | 90.10 | 75.20 | 90.03 | 92.33 | **93.14** |
| Wellcome | 85.94 | 88.09 | 92.67 | 93.22 | **93.39** |
| **mAP** | 81.59 | 85.94 | 91.56 | 92.89 | **93.33** |

**TABLE 4.** Comparison of results from state-of-the-art detection & recognition methods and Proposed Method I on the BSVSO dataset.

| Classes | SSD | YOLO v2 | YOLO v3 | YOLO v4 | Our proposed method I |
|---|---|---|---|---|---|
| Signboard | 52.73 | 65.22 | 68.14 | 68.55 | **70.01** |
| **mAP** | 52.73 | 65.22 | 68.14 | 68.55 | **70.01** |

**TABLE 5.** Comparison of results from Proposed Methods I and II on our datasets.

| Classes | Our proposed method I | Our proposed method II |
|---|---|---|
| Carrefour | 88.19 | **89.30** |
| Domino | 97.08 | **97.43** |
| Family Mart | 96.71 | **97.36** |
| Gas | 96.10 | **96.30** |
| Hi-Life | 95.34 | **95.55** |
| KFC | 92.16 | **92.28** |
| McDonald | 91.12 | **91.24** |
| Mos Burger | 90.46 | **91.69** |
| Ok Mart | 93.34 | **93.99** |
| Post | 91.83 | **94.65** |
| PxMart | 93.08 | **93.54** |
| 7-Eleven | 94.68 | **95.70** |
| Starbucks | 93.14 | **93.86** |
| Wellcome | 93.39 | **95.50** |
| **mAP** | 93.33 | **94.17** |

**TABLE 6.** Comparison of results from Proposed methods I and II on BSVSO datasets.

| Classes | Our proposed method I | Our proposed method II |
|---|---|---|
| Signboard | 70.01 | **70.92** |
| **mAP** | 70.01 | **70.92** |

**TABLE 7.** Different model configurations.

| Configurations | Domain adaptation | Non-local blocks | Attention mechanism |
|---|---|---|---|
| I | × | × | × |
| II | ✓ | × | × |
| III | × | ✓ | × |
| IV | × | × | ✓ |
| V | ✓ | ✓ | × |
| VI | ✓ | × | ✓ |
| VII | × | ✓ | ✓ |
| VIII | ✓ | ✓ | ✓ |

588 images as target dataset. The results for the pre-trained model using domain adaptation are listed in Table 2.

In Table 2, the label classification results in the source data are almost entirely above 70% in epochs 500 to 5,000. Meanwhile, the highest and the lowest results of the target are 60.49% and 41.28%, respectively. For the pre-training model, we took the weight in epoch 4,000 since the label classification of the target achieved better results than others did. In the domain adaptation for classification, the essential accuracy value is the target accuracy. Label classification accuracy caters to the proper categorization of images in each domain (For example, classifying an image into labels such as Okmart, Familymart, etc.). On the other hand, domain discrimination accuracy aids in discriminating whether an image belongs to the source domain or the target domain. The discriminator in the model takes care of the domain discrimination, while the classifier performs the label classification of the images.

### D. COMPARISON OF THE ATTENTION NETWORK RESULT WITH COMMON PRE-TRAINED MODEL

SSD, YOLOv2, YOLOv3 and YOLOv4 are considered to be the state-of-the-art deep learning methods, especially in detection and recognition. Our proposed method was tested on our dataset and in BSVSO since both datasets have annotation files and similar objects for the experiment. The comparison results between the detection and recognition method in our dataset and BSVSO can be seen in Tables 3 and 4, respectively.

**TABLE 8.** Influence of domain adaptation, non-local blocks, and attention mechanism using our dataset.

| Classes | I<br>mAP | II<br>mAP | III<br>mAP | IV<br>mAP | V<br>mAP | VI<br>mAP | VII<br>mAP | VIII<br>mAP |
|---|---|---|---|---|---|---|---|---|
| Carrefour | 62.74 | 70.72 | 69.96 | 80.65 | 84.95 | 85.05 | 88.19 | **89.30** |
| Domino | 90.17 | 86.15 | 91.04 | 91.06 | 94.31 | 96.61 | 97.08 | **97.43** |
| Family Mart | 91.23 | 88.09 | 92.76 | 86.41 | 93.70 | 96.15 | 96.71 | **97.36** |
| Gas | 78 | 90.50 | 93.09 | 93.67 | 95.85 | 96.66 | 96.10 | **96.30** |
| Hi-Life | 76 | 87.80 | 92.20 | 86.59 | 93.76 | 94.41 | 95.34 | **95.55** |
| KFC | 87.93 | 75.82 | 79.10 | 85.81 | 88.68 | 89.17 | 92.16 | **92.28** |
| McDonald | 78.57 | 68.51 | 83.64 | 80.55 | 86.76 | 86.87 | 91.12 | **91.24** |
| Mos Burger | 66.67 | 74.87 | 82.00 | 82.14 | 87.50 | 88.13 | 90.46 | **91.69** |
| Ok Mart | 92 | 85.03 | 88.56 | 88.08 | 87.39 | 90.22 | 93.34 | **93.99** |
| Post | 76 | 88.77 | 83.72 | 90.15 | 91.77 | 94.03 | 91.83 | **94.65** |
| PxMart | 86 | 88.24 | 82.69 | 89.64 | 91.21 | 93.08 | 93.08 | **93.54** |
| 7-Eleven | 80.93 | 86.09 | 89.51 | 85.07 | 91.46 | 93.15 | 94.68 | **95.70** |
| Starbucks | 90.1 | 69.66 | 74.82 | 75.20 | 85.51 | 89.02 | 93.14 | **93.86** |
| Wellcome | 85.94 | 83.95 | 89.57 | 88.09 | 93.17 | 93.28 | 93.39 | **95.50** |
| **mAP** | **81.59** | **81.73** | **85.19** | **85.94** | **90.43** | **91.84** | **93.33** | **94.17** |

**TABLE 9.** Influence of domain adaptation, non-local blocks, and attention mechanism using the BSVSO dataset.

| Classes | I<br>mAP | II<br>MAP | III<br>mAP | IV<br>mAP | V<br>mAP | VI<br>mAP | VII<br>mAP | VIII<br>mAP |
|---|---|---|---|---|---|---|---|---|
| Signboard | 62.55 | 63.67 | 64.18 | 66.96 | 67.51 | 68.33 | 70.01 | **70.92** |
| **mAP** | **62.55** | **63.67** | **64.18** | **66.96** | **67.51** | **68.33** | **70.01** | **70.92** |

As shown in Table 3, we achieved a higher mean average precision (mAP) than the other methods did. This means that our Proposed method I is able to recognize and detect the signboard objects. Moreover, our proposed method achieves higher accuracy by one class, as shown in Table 4, and can be used in one or more categories.

### E. COMPARISON OF PRETRAINED MODEL IN DOMAIN ADAPTATION

We compared the common pre-trained models from public datasets such as Imagenet, MS COCO and Pascal VOC with our pre-trained model using the domain adaptation method. Likewise, the proposed method was tested on our dataset and the BSVSO dataset due to the aforementioned reasons. The comparison of results between the detection and recognition methods in our dataset and the BSVSO dataset can be seen in Tables 5 and 6, respectively.

In Tables 5 and 6, we can see that our dataset contains 14 classes reaching mAP above 94.17%, with 0.84% improvement. For the BSVSO dataset, we also applied our proposed method, which has the highest accuracy of 70.92% with a 0.91% improvement. The comparison results for SSD, YOLOv2, YOLOv3, YOLOv4, and our proposed method are shown in Figure 9.

In Figure 8, SSD, YOLOv2, YOLOv3 and YOLOv4 have the worst results on both our dataset and the BSVSO dataset, which has one class category. In the SSD results, misclassified objects include not only signboards, but also other objects, such as walls. Moreover, YOLOv2 and YOLOv4 are still unstable for detection and recognition compared to our proposed method for the BSVSO datasets. On our dataset, the proposed method II could detect logos accurately, but the proposed method I sometimes has some redundant bounding boxes. Meanwhile, on the BSVSO dataset, our proposed method proved to be more stable and effective than YOLOv4 in the results.

### F. ABBLATION STUDY

We performed an ablation study on the overall framework, including domain adaptation, non-local blocks, and attention mechanism.

In Table 7, we analyze eight configurations to identify the role of each module in our network. The comparison of performance on our dataset and the BSVSO dataset is presented in Tables 8 and 9, respectively.

In Table 8 and Table 9, Configuration V can reach 90.43% and 67.51% with domain adaptation and non-local blocks. The results signify that domain adaptation and non-local blocks can greatly improve the baseline performance of the network.

## V. CONCLUSION

This paper solves detection and recognition issues on logo images by integrating a domain adaptation as the pre-trained model and an attention network as the detection and recognition system for deep learning methods. Domain adaptation is a useful method since it can reduce the loss function between the source and the target. It is adapted from the GAN using encoder-decoder architecture and exhibits improvement, especially for logo detection and recognition. Both our detection and recognition networks combine nonlocal networks and channel attention to determine the important feature in the logo images. The proposed methods achieved higher accuracy compared to the state-of-the-art detection and recognition methods on our datasets. Although new detection and recognition methods (e.g., YOLOv4) have already been released, our method achieves the best results in the detection and recognition of logo images.

## REFERENCES

[1] W. Rahmaniar and W.-J. Wang, "Real-time automated segmentation and classification of calcaneal fractures in CT images," *Appl. Sci.*, vol. 9, no. 15, p. 3011, Jul. 2019.

[2] I. Fehervari and S. Appalaraju, "Scalable logo recognition using proxies," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 715–725.

[3] O. Orti, R. Tous, M. Gomez, J. Poveda, L. Cruz, and O. Wust, "Real-time logo detection in brand-related social media images," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, in Lecture Notes in Computer Science: Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11507, 2019, pp. 125–136.

[4] D. M. Montserrat, Q. Lin, J. Allebach, and E. Delp, "Scalable logo detection and recognition with minimal labeling," in *Proc. IEEE 1st Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 152–157.

[5] C. Liao, W. Wang, K. Sakurada, and N. Kawaguchi, "Image-matching based identification of store signage using web-crawled information," *IEEE Access*, vol. 6, pp. 45590–45605, 2018.

[6] Q. Yu, C. Szegedy, M. C. Stumpe, L. Yatziv, V. Shet, J. Ibarz, and S. Arnoud, "Large scale business discovery from street level imagery," 2015, *arXiv:1512.05430*. [Online]. Available: https://arxiv.org/abs/1512.05430

[7] C.-Y. Hong, C.-Y. Lin, and T. K. Shih, "Automatic signboard detection and semi-automatic ground truth generation," in *Proc. 12th Int. Conf. Ubi-Media Comput. (Ubi-Media)*, Aug. 2019, pp. 256–261.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9905, 2016, pp. 21–37.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[10] Z. Huang, Z. Zhong, L. Sun, and Q. Huo, "Mask R-CNN with pyramid attention network for scene text detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 764–772.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[12] M. S. I. Toaha, C. R. Rahman, S. BinAsad, T. Ahmed, M. A. Proma, and S. M. S. Haque, "Automatic signboard detection from natural scene image in context of Bangladesh Google street view," 2020, *arXiv:2003.01936v2*. [Online]. Available: https://arxiv.org/abs/2003.01936v2

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[15] W. Rahmaniar, W. J. Wang, and H. C. Chen, "Real-time detection and recognition of multiple moving objects for aerial surveillance," *Electronics*, vol. 8, no. 12, pp. 1–17, 2019.

[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: https://arxiv.org/abs/2004.10934

[18] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[19] C.-T. Lin, "Cross domain adaptation for on-road object detection using multimodal structure-consistent image-to-image translation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3029–3030.

[20] Z. He, B. Yang, C. Chen, Q. Mu, and Z. Li, "CLDA: An adversarial unsupervised domain adaptation method with classifier-level adaptation," *Multimedia Tools Appl.*, vol. 79, nos. 45–46, pp. 33973–33991, Dec. 2020.

[21] W. Teng, N. Wang, H. Shi, Y. Liu, and J. Wang, "Classifier-constrained deep adversarial domain adaptation for cross-domain semisupervised classification in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 789–793, May 2020.

[22] Q. Wang, Z. Li, Q. Zou, L. Zhao, and S. Wang, "Deep domain adaptation with differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3093–3106, 2020.

[23] X. Jia and F. Sun, "Unsupervised deep domain adaptation based on weighted adversarial network," *IEEE Access*, vol. 8, pp. 64020–64027, 2020.

[24] G. Shojaei and F. Razzazi, "Semi-supervised domain adaptation for pedestrian detection in video surveillance based on maximum independence assumption," *Int. J. Multimedia Inf. Retr.*, vol. 8, no. 4, pp. 241–252, Dec. 2019.

[25] W. Li, M. Wang, H. Wang, and Y. Zhang, "Object detection based on semi-supervised domain adaptation for imbalanced domain resources," *Mach. Vis. Appl.*, vol. 31, no. 3, pp. 1–18, Mar. 2020.

[26] H. N. Oliveira, E. Ferreira, and J. A. D. Santos, "Truly generalizable radiograph segmentation with conditional domain adaptation," *IEEE Access*, vol. 8, pp. 84037–84062, 2020.

[27] Y. Li, W. Hu, H. Li, H. Dong, B. Zhang, and Q. Tian, "Aligning discriminative and representative features: An unsupervised domain adaptation method for building damage assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 6110–6122, 2020.

[28] X. Gu, Y. Guo, F. Deligianni, and G.-Z. Yang, "Coupled real-synthetic domain adaptation for real-world deep depth enhancement," *IEEE Trans. Image Process.*, vol. 29, pp. 6343–6356, 2020.

[29] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 28, 2020, doi: 10.1109/TPAMI.2020.2991050.

[30] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6103–6115, Dec. 2019, doi: 10.1109/TIP.2019.2924174.

[31] Z. Gao, L. Guo, W. Guan, A.-A. Liu, T. Ren, and S. Chen, "A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-R2," *IEEE Trans. Image Process.*, vol. 30, pp. 767–782, 2021, doi: 10.1109/TIP.2020.3038372.

[32] Z. Gao, L. Guo, T. Ren, A.-A. Liu, Z.-Y. Cheng, and S. Chen, "Pairwise two-stream ConvNets for cross-domain action recognition with small data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 9, 2021, doi: 10.1109/TNNLS.2020.3041018.

[33] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 1234–1241.

[34] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11211, 2018, pp. 3–19.

[35] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[36] X. Xiang, Y. Zhang, and A. El Saddik, "Pavement crack detection network based on pyramid structure and attention mechanism," *IET Image Process.*, vol. 14, no. 8, pp. 1580–1586, Jun. 2020.

[37] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.

[38] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7127–7136.

[39] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.

[40] S. Zhang, H. Li, and W. Kong, "Object counting method based on dual attention network," *IET Image Process.*, vol. 14, no. 8, pp. 1621–1627, Jun. 2020.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[42] R. Zhang, M. Yang, X. Bai, B. Shi, D. Karatzas, S. Lu, C. V. Jawahar, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, and M. Liao, "ICDAR 2019 robust reading challenge on reading Chinese text on signboard," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1577–1581.

**ERVIN YOHANNES** received the bachelor's degree from the Department of Informatics Engineering, University of Brawijaya, Malang, Indonesia, in 2013, the dual master's degrees from the Department of CSIE, National Central University (NCU), Taoyuan City, Taiwan, and the Department of Computer Science, University of Brawijaya, in 2017. He is currently pursuing the Ph.D. degree with the Department of CSIE, NCU. His research interests include object detection and recognition, context-aware, computer vision, human–computer interaction, and deep learning.

**CHIH-YANG LIN** (Senior Member, IEEE) was with the Advanced Technology Center, Industrial Technology Research Institute of Taiwan, Hsinchu, Taiwan, from 2007 to 2009. He was a Postdoctoral Fellow with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, in 2009. He joined Asia University, Taichung, Taiwan, in 2010, where he is an Assistant Professor and then became an Associate Professor, in 2013, and a Professor, in 2016. He was the Chair of the Department of Bioinformatics and Medical Engineering, Asia University, from August 2014 to January 2017. He is currently the Deputy Chief of Global Affairs Office and a Professor with the Department of Electrical Engineering, Yuan-Ze University, Taoyuan City, Taiwan. He has published over 100 papers in international conferences and journals with more than 1300 citations. His research interests include computer vision, machine learning, deep learning, image processing, big data analysis, and the design of surveillance systems. He received the Best Paper Award from Pacific-Rim Conference on Multimedia (PCM), in 2008, the Best Paper Award and Excellent Paper Award from Computer Vision, Graphics and Image Processing Conference, in 2009, 2013, and 2019, and the Best Paper Award from the 6th International Visual Informatics Conference 2019 (IVIC'19). He has served as a Session Chair and the Publication Chair Workshop Organizer for many international conferences, including *AHFE*, ICCE, ACCV, IEEE Multimedia Big Data, ACM IH&MMSec, APSIPA, and *CVGIP*. He is a Regular Reviewer of the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Multimedia, IEEE Access, and many other Elsevier journals.

**TIMOTHY K. SHIH** (Senior Member, IEEE) currently a Distinguished Professor and the Vice Dean of the College of EECS, National Central University, Taiwan. He is also the Director of the Innovative AI Research Center. His current research interests include multimedia computing, human–computer interaction, and distance learning. He is a Fellow of the Institution of Engineering and Technology (IET). In addition, he is a Senior Member of ACM. He is currently the Associate Editor of the IEEE Computing. Before that, he was the Associate Editor of the IEEE Transactions on Learning Technologies, *ACM Transactions on Internet Technology*, and the IEEE Transactions on Multimedia.

**CHEN-YA HONG** received the B.S. degree from the National University of Kaohsiung (NUK), Kaohsiung, Taiwan, and the M.S. degree from the Department of Computer Science and Information Engineering, National Central University (NCU), Taoyuan City, Taiwan. Her research interests include computer vision, image processing, and deep learning.

**AVIRMED ENKHBAT** received the B.S. degree in computer science and the M.S. degree in applied sciences and engineering from the National University of Mongolia, Mongolia, in 2011and 2016, respectively. He is currently pursuing the Ph.D. degree in computer science and information engineering from National Central University (NCU), Taoyuan City, Taiwan. His research interests include computer vision, human–computer interaction, and gesture recognition.

**FITRI UTAMININGRUM** was born in Surabaya, East Java, Indonesia. She received the bachelor's degree in electrical engineering from the National Institute of Technology, the master's degree in electrical engineering from Brawijaya University, Malang, Indonesia, and the Doctor of Engineering degree in computer science and electrical engineering from Kumamoto University, Japan. She is currently an Associate Professor with the Faculty of Computer Science, Brawijaya University, Indonesia. She is a Coordinator with the Computer Vision Research Group and a full-time Lecturer with Brawijaya University, Indonesia. She has been published her worked in several reputable journals and conferences indexed by Scopus. Her research interests include smart wheelchair, especially on the development of image algorithms.