IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 12, 2020, accepted October 20, 2020, date of publication October 23, 2020, date of current version November 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033289

A Deep Learning-Based Hybrid Framework for Object Detection and Recognition in Autonomous Driving

YANFEN LI[1], HANXIANG WANG[1], L. MINH DANG[1], TAN N. NGUYEN[2], DONGIL HAN[1], (Member, IEEE), AHYUN LEE[3], INSUNG JANG[3], AND HYEONJOON MOON[1]

[1]Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
[2]Department of Architectural Engineering, Sejong University, Seoul 05006, South Korea
[3]City & Geospatial ICT Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea

Corresponding author: Hyeonjoon Moon (hmoon@sejong.ac.kr)

This work was supported by Electronics and Telecommunications Research Institute (ETRI) Grant funded by the Korean Government [20ZR1200, DNA-Based National Intelligence Core Technology Development] and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540).

ABSTRACT As a key technology of intelligent transportation system, the intelligent vehicle is the carrier of comprehensive integration of many technologies. Although vision-based autonomous driving has shown excellent prospects, there is still a problem of how to analyze the complicated traffic situation by the collected data. Recently, autonomous driving has been formulated as many tasks separately by using different models, such as object detection task and intention recognition task. In this study, a vision-based system was developed to detect and identity various objects and predict the intention of pedestrians in the traffic scene. The main contributions of this research are (1) an optimized model was presented to detect 10 kinds of objects based on the structure of YOLOv4; (2) a fine-tuned Part Affinity Fields approach was proposed to estimate the pose of pedestrians; (3) Explainable Artificial Intelligence (XAI) technology is added to explain and assist the estimation results in the risk assessment phase; (4) an elaborate self-driving dataset that includes several different subsets for each corresponding task was introduced; and (5) an end-to-end system containing multiple models with high accuracy was developed. Experimental results proved that the total parameters of optimized YOLOv4 are reduced by 74%, which satisfies the real-time capability. In addition, the detection precision of the optimized YOLOv4 achieved an improvement of 2.6% compared to the state-of-the-art.

INDEX TERMS Deep learning, intention recognition, object detection, risk assessment.

I. INTRODUCTION

Rapid urbanization has highlighted a series of problems, especially in the aspect of transportation, which severely limits travel and has certain security risks. Even though some progress has been made in the existing object detection technologies in self-driving, there still exist potential risk factors of collision as motor cars are surrounded by many objects in daily life, including some uncontrollable moving objects (pedestrians and vehicles) and static objects (traffic lights and signs). Therefore, it is necessary to promptly detect various static objects and accurately estimate the intention of moving objects.

The associate editor coordinating the review of this manuscript and approving it for publication was Fan-Hsun Tseng.

In the object detection tasks, the main deep learning methods are divided as one-stage detection algorithms and two-stages detection algorithms. YOLO [1] and SSD [20] are one-stage detection methods that directly convert the detection problem to a unified regression problem. Due to the characteristics of the structure, the one-stage methods are faster than two-stage methods. Faster R-CNN [3] is a typical two-stage network that generates a series of candidate bounding boxes and then classifies each object by using the Convolutional Neural Network (CNN). From the aspects of detection and localization precision, the two-stage methods perform better than most of the one-stage methods. In this study, the proposed model with multiple tasks is based on the one-stage methods to reduce the time used for the object detection phase.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

In the intention recognition section, most researchers rely on appearance-based features [6] and skeleton-based features [5], [7]. The intention of pedestrians is recognized by using the pose estimation algorithm in this study. For the dangerous vehicle assessment stage, the directions of vehicles are predicted by different references, such as the taillight of vehicles [9] and traffic features [8]. Besides, the traffic light is an important sign in the traffic scene, which is recognized by using a CNN model after filtering candidate traffic lights [27].

The main contributions of this paper are as followsčž

- In the object detection phase, the structure is optimized based on YOLOv4 [18] model, and the computation complexity is significantly reduced while improving the detection accuracy.
- In the pose estimation section, the Part Affinity Fields method is fine-tuned to improve the inference time.
- Explainable Artificial Intelligence (XAI) technology is added to explain and assist the estimation results in the risk assessment phase.
- Three datasets for different tasks are collected and validated manually.
- An end-to-end system with high accuracy is proposed to integrate several phases, including object detection, pose estimation, intention recognition, dangerous vehicle recognition, and traffic light recognition.

The rest of the paper is arranged as follows: Section 2 presents a literature review on previous methods. The description of the different methods used in this research is discussed in Section 3. Detailed information and acquisition process of datasets are illustrated in Section 4. In Section 5, the obtained results for each task and comprehensive performance of the proposed framework are analyzed. Finally, the conclusion and future study are stated in Section 6.

II. RELATED WORK

A. OBJECT DETECTION

Recently, multi-objects detection has been a prevailing topic that attracting a lot of researchers in the field of autonomous driving. A one-stage method You Only Look Once (YOLO) was first presented to address object detection as a regression problem. As one of the state-of-the-art works, YOLO can achieve robust and fast performance in object detection, but the spatial constraint of the model limits the predicted amount of objects [1]. Another one-stage method is named the single shot Multi-box detection (SSD) [20]. For a 300×300 input size, the SSD model can achieve at 59 FPS, and 74.3% mean Average Precision (mAP) on the PASCAL VOC dataset, which is greatly superior to the real-time YOLO [1]. Besides, a unified network was performed for object detection. Experiments show the processing speed of the method is slower than the YOLO [1], but it achieved better performance in mAP due to the improved gripping process [2]. Compared with most of the one-stage methods, the two-stage methods can obtain more accurate detection results, but the detection speed is slower. For example, Faster R-CNN is a kind of two-stage method, which optimized the overall accuracy by introducing a region proposal network [3]. In previous work, the performance of the autonomous driving dataset BDD100K has not been reported. Thus, the latest method YOLOv4 [18] is superior to the other state-of-art detectors and was optimized and tested on BDD100K dataset in this study.

B. INTENTION RECOGNITION

Since deep learning has significantly enhanced object detection performance, some extensions have been proposed to estimate the postures of pedestrians and vehicles. Based on the appearances of pedestrians, a CNN model was proposed to classify the pedestrians' head pose and body orientation, and the method is available for still images and image sequences [6]. In another work, a neural network using appearance features was provided to predict the location and keypoints in pose estimation [28]. In contrast to CNN-appearance-based methods, the dynamical model with Gaussian processes was presented to predict paths and poses of pedestrians by analyzing fitted skeletons [7]. The proposed skeleton-based intention recognition was compared with the appearance-based model to evaluate the effectiveness, and the results proved that the former achieved better performance [5]. However, classification accuracy on the skeleton features obtained 88% by using Random Forest algorithm, which is not satisfactory in the self-driving system.

C. RISK ASSESSMENT

For risk assessment, the recognition of traffic lights and the moving trend for vehicles are important factors to avoid traffic accidents. A recurrent network with effective performance was used to predict the intention of drivers at different types of intersections [8]. In another work, an end-to-end method combining CNN and Long Short-Term Memory (LSTM) was applied to recognize the direction of vehicles based on vehicle taillights [9]. As the key signal on the road, the detected traffic lights were recognized by a CNN model after filtering the traffic light candidates by the importance map in a real-time system [27]. A multi-task learning approach combining object detection and distance estimation was presented to probe the characteristics of dangerous objects with different distances [4], which achieved a better performance of 2.27% than SSD method on the KITTI dataset. Even though there are some existing methods for safe driving, the method that can jointly process object detection, intention recognition, and risk assessment is not considered in the previous work.

There are two main categories considered in self-driving, including still objects and dynamic objects. Herein, a vision-based model with multiple tasks was proposed to detect various objects and assess the posture of pedestrians and vehicles (dynamic objects) in this study. Besides, the traffic lights (still objects) are recognized to indicate whether the automatic driving system should continue to drive.

III. METHODOLOGY

In this part, the whole diagram of the proposed framework is shown in Fig. 1. In real traffic scenes, the risk factors

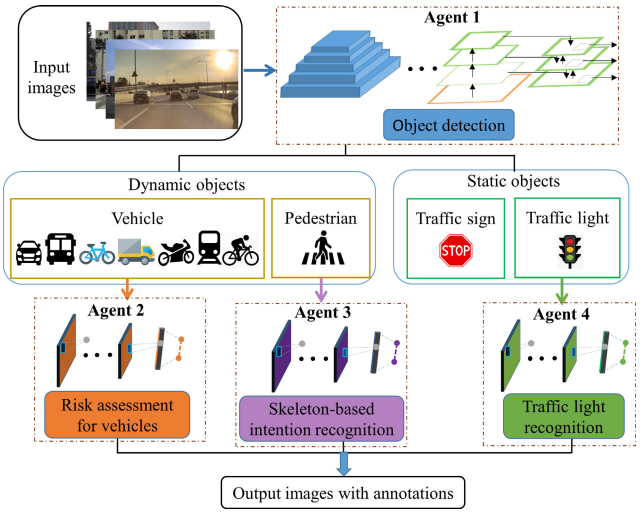


FIGURE 1. Flowchart of the proposed framework. There are four tasks in the proposed framework, including object detection, risk assessment for vehicles, skeleton-based intention recognition, and traffic light recognition.

affecting self-driving mainly include dynamic objects (pedestrians and different kinds of vehicles) and static objects (traffic lights and stop signs). Thus, various objects are localized and detected using an optimized YOLOv4 model in the first step. And then, based on the detected objects, it is extremely necessary to estimate the intentions of vulnerable road users for performing safe self-driving without traffic accidents. Furthermore, the recognition of traffic lights is an essential part of unmanned driving, and its recognition accuracy is directly related to the safety of intelligent driving. In this paper, the model used for object detection is described in detail (see in Section A). After that, the skeleton-based intention recognition for pedestrians is illustrated in Section B. Finally, the recognition of dangerous vehicles and traffic lights with XAI technology is explained in Section C.

A. OBJECT DETECTION

Object detection is the main task in the whole system, which gives the bounding boxes and categories probabilities for each object. As the state-of-art object detector, YOLOv4 obtained better performance in detection speed (FPS) and detection accuracy (mAP) than all available methods on MS COCO dataset [22]. The model structure of YOLOv4 is composed of CSPDarknet-53, Spatial Pyramid Pooling in Deep Convolutional networks (SPPnet), Path Aggregation Network (PANet), and three YOLO heads, as shown in Fig. 2. As the backbone of YOLOv4, CSPDarknet-53 is responsible for extracting deep features of the input image through 5 Resblock bodies (C1-C5). The network contains 53 convolution layers with the sizes of 1×1 and 3×3, and each convolution layer is connected with a batch normalization (BN) layer and a Mish activation layer. Furthermore, all activation functions in YOLOv4 are replaced with leaky-ReLU that requires less computation. SPPnet effectively increased the receptive field of the model through different max-pooling layers with the

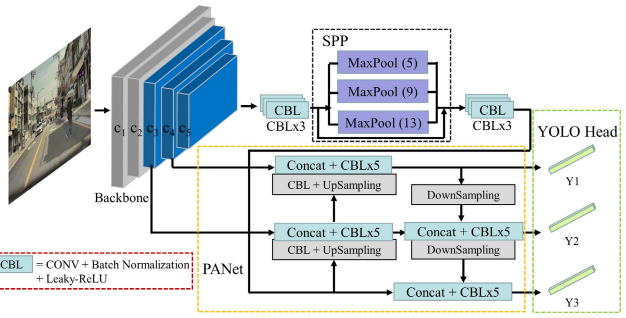


FIGURE 2. Overall structure of YOLOv4, including CSPDarknet (backbone), SPPnet, PANet and 3 YOLO heads. The modified part (C3-C5) is in the backbone of the model, which is highlighted in blue color.

size of 5, 9, and 13. and PANet used top-down and bottom-up approaches to extract features repeatedly. Three YOLO heads with sizes of 19×19, 38×38, and 76×76 are used to fuse and interact with feature maps of different scales to detect objects of different sizes. In this study, the original model structure is modified to reduce the computation complexity by using the layer pruning method. As shown in Fig. 2, eight shortcut structures in the backbone structure were removed from C3-C5 Resblock bodies highlighted in blue color.

The loss function used in the YOLOv4 model consists of three parts: object localization offset loss $L_{conf}\mathrm{L_{loc}}$, object confidence loss L_{conf} and object classification loss L_{cla}, as shown in (1), where λ is the balance coefficient.

$$Loss = \lambda_1 L_{conf} + \lambda_2 L_{cla} + \lambda_3 L_{loc} \tag{1}$$

$$L_{conf} = -\sum (Obj_i \ln(p_i) + (1 - Obj_i)\ln(1 - p_i)) \tag{2}$$

$$L_{cla} = -\sum_{i \epsilon Box} \sum_{j \epsilon class} (O_{ij} \ln(p_{ij}) + (1 - O_{ij})\ln(1 - p_{ij})) \tag{3}$$

$$L_{loc} = 1 - IOU(A, B) + \frac{d^2_{AB}(A_{ctr}, B_{ctr})}{l^2} + \alpha v \tag{4}$$

In (2), Obj_i indicates whether there is an object in the predicted object bounding box i, and the result value is 0 or 1. p_i refers to the probability that there is a real object in the prediction box. The probability value is obtained by calculating the sigmoid function. In (3), O_{ij} and p_{ij} mean whether there is a j-class object and probability in the prediction boundary box i, respectively. YOLOv4 adopts the Complete Intersection Over Union (CIOU) algorithm [23] to calculate the object localization offset loss, as shown in (4), where the aspect ratio αv and Euclidean distance of the center point (A_{ctr}, B_{ctr}) for the predicted bounding box A and the GroundTruth bounding box B are calculated.

In this study, the original YOLOv4 model is mainly modified from the aspect of network structure. First, the network structure of YOLOv4 enhances the learning of small objects, but the detection performance for some large objects is not good in real testing. That is because the weight of high dimensional features becomes lower after the fusion of high dimensional features and low dimensional features. Based on the above analysis of the confidence loss formulas, the confidence weight for each YOLO head were designed to

Type	Filters before	Filters after	Size	Output	Shortcuts before	Shortcuts after
Convolutional	32	27	3x3	608x608		
Convolutional	64	64	3x3/2	304x304		
Convolutional	64	62	1x1			
Convolutional	32	32	1x1			
Convolutional	64	64	3x3		x 1	x 1
Residual				304x304		
Convolutional	64	64	1x1			
Residual						
Convolutional	64	63	1x1			
Convolutional	128	99	3x3/2	152x152		
Convolutional	64	61	1x1			
Convolutional	64	41	1x1			
Convolutional	64	64	3x3		x 2	x 2
Residual				152x152		
Convolutional	64	41	1x1			
Residual						
Convolutional	128	118	1x1			
Convolutional	256	189	3x3/2	76x76		
Convolutional	182	92	1x1			
Convolutional	128	87	1x1			
Convolutional	128	120	3x3		x 8	x 5
Residual				76x76		
Convolutional	256	217	1x1			
Residual						
Convolutional	128	118	1x1			
Convolutional	512	365	3x3/2	38x38		
Convolutional	256	235	1x1			
Convolutional	256	171	1x1			
Convolutional	256	256	3x3		x 8	x 5
Residual				38x38		
Convolutional	256	228	1x1			
Residual						
Convolutional	512	460	1x1			
Convolutional	1024	676	3x3/2	19x19		
Convolutional	512	400	1x1			
Convolutional	512	286	1x1			
Convolutional	512	512	3x3		x 4	x 2
Residual				19x19		
Convolutional	512	373	1x1			
Residual						
Convolutional	1024	510	1x1			

The parameters for layers and channels before pruning.
The parameters for layers and channels after pruning.
The shortcut structures.

FIGURE 3. Modifications of CSPDarknet.

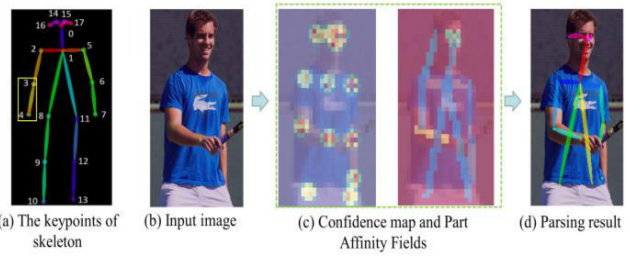


FIGURE 4. Overall process of PAFs. A total of 18 human keypoints are detected for pose estimation, as shown in (a). (b) and (d) represent the input image and output image respectively. In (c), the confidence map and part affinity fields generated from two-branch CNN are used to predict body part locations and parts association.

keep balance. Secondly, compared with YOLOv3, the mAP of YOLOv4 is increased by nearly 8%, but the inference speed is not significantly improved. Thus, the channel and layer pruning algorithm [24] is adopted to reduce the model parameters. Fig. 3 presents the modification of the backbone structure of YOLOv4. In this experiment, the channel pruning method reduces the parameters of the whole model by 66%, from 64,363,101 to 21,749,380. Besides, 8 shortcut structures were removed by the layer pruning method using in the backbone part. The total parameters of the final simplified model are only 26% of the original YOLOv4. According to the testing result, the inference time of the modified model obtained 0.012s faster than the original one.

Moreover, the BDD100K dataset is used in the training phase of the object detection task, and the number of each class has a large gap. For example, there are 714,121 car samples and only 136 train samples in the training set. Although the original YOLOv4 adopted some techniques to solve the imbalanced data problem, such as MixUp, CutMix, Mosaic, and Focal loss algorithm, the imbalanced data problem still exists according to the testing results. To further improve the performance of YOLOv4 on detection accuracy, images with a small sample size, including bike, motor, train, and rider, are selected from the training set for intensive training. The selected images contain not only the samples of a specific category but also may include the samples of other categories. Besides, this step is performed after the training of each epoch. Therefore, the model can train samples from all categories to avoid biased training.

B. INTENTION RECOGNITION

Based on the monocular pedestrian detection, the method of using pose estimation as the main information to predict whether pedestrians have the intention of crossing the road is explored in this section. As a part of the proposed system, the intention recognition is combined with the object detection algorithm to predict the intention of detected pedestrians. Firstly, the area of the detected pedestrian is obtained by YOLOv4 and input into the method called Part Affinity Fields (PAFs) [17] to generate human skeleton features. In order to avoid the situation that the object detection algorithm does not capture the required human features, the area of the bounding box is slightly expanded by increasing the original side length by 1/6. (The same method is used in the process of risk assessment). Secondly, the CNN model is used to analyze the image with skeleton features overlayed and identify the intention of pedestrians. Finally, the recognition results are returned to the YOLOv4 model and displayed as a label on the bounding box.

1) POSE ESTIMATION

In the pose estimation stage, the PAFs model is adopted to acquire the skeleton features. The original structure of PAFs composes of the first 10 layers of VGG-19 [12] and six groups of CNN structures containing two branches. The two branches are responsible for generating confidence maps and PAFs, respectively. As shown in Fig. 4 (a), there are 18 key points of human body parts, including nose, neck, right shoulder, left shoulder, right elbow, left elbow, right twist, left twist, right hip, left hip, right knee, left knee, right ankle, left ankle, right eye, left eye, right ear, and left ear [30]. Based on the bottom-up approach, this algorithm predicts the part locations and the connection relations of these parts in the input image Fig. 4 (b), and generates the confidence maps and part affinity fields through multiple learning processes, as shown in Fig. 4 (c). Finally, confidence maps and part affinity fields are assembled to generate the parsing result Fig. 4 (d).

Though the experiments, it is found that when the human area is very small, the original PAFs method misses some body parts, such as ankles and wrists, so more detailed features need to be extracted in the training phase to improve the location precision. Moreover, with the increase of the number of stages, the precision is not significantly improved, but the computational complexity of the model is increased. Considering the above two problems, the original structure of PAFs model is modified, as shown in Fig. 5. Firstly, the number of convolution layers for feature extraction is increased from 10 to 14, and BN layer is added to the model to achieve the decoupling between layers, so as to accelerate the training speed of the model. Secondly, 6 groups of CNN structures used to generate confidence maps, and PAFs were reduced

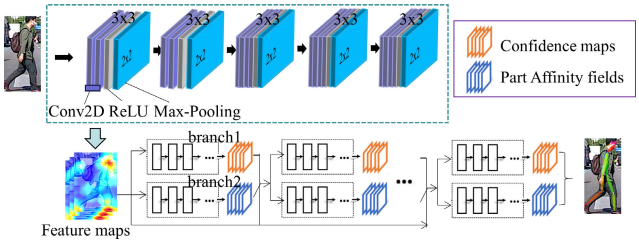


FIGURE 5. Architecture of the modified PAFs. The feature maps extracted from 14 convolutional layers in the blue box are input to two-branch CNN structure to generate confidence maps and part affinity fields, which are used to obtain keypoints and associate body parts.

to 4 stages to simplify the structure of the model. Finally, the 7×7 convolution kernel in the CNN structure is replaced by multiple 3×3 convolution kernels, which can enhance the network capacity and reduce the number of parameters.

2) INTENTION RECOGNITION

After the pose estimation stage, the skeleton features are acquired and fed into a CNN model to recognize the intention of pedestrians. A suitable CNN model is selected by comparing the recognition accuracy and recognition speed. Besides, the performances of skeleton-based features and appearance-based features are evaluated and compared in the experiment result section.

C. CNN-BASED RISK ASSESSMENT

The risk assessment section is considered as two CNN-based classification tasks, dangerous vehicle estimation, and traffic light recognition. Due to the special structure of CNN, it has unique advantages in various tasks, such as image classification, object detection, and segmentation. Different from common neural networks, the basic structure of CNN includes convolutional layers and pooling layers, and the essence is feature extraction and parameter reduction. CNN can directly learn the mapping relations between a large number of inputs and outputs without any precise mathematical expression. Besides, the complexity of the CNN model is reduced greatly through three strategies (local receptive field, weight sharing, and down sampling).

In this study, five different CNN models are evaluated and compared to achieve high recognition accuracy. VGGNet demonstrates that the depth of the network structure is the essential part of the performance for an algorithm. By repeatedly stacking 3×3 small convolution kernel and 2×2 max-pooling layer, convolution neural network with 16-19 layers depth is successfully constructed [12]. GoogLeNet has made a bold attempt in the network structure. Although the architecture of GoogLeNet has 22 layers, it is much smaller compared with AlexNet and VGGNet. Thus, GoogLeNet is more suitable when computing resources or memory are restricted [13]. As a residual network, Resnet can maintain strong accuracy growth with depth increasing, which effectively avoids the problem that the accuracy of VGG model decreases with the increase of number of layers [14].

FIGURE 6. Sample images from the four different datasets.

TABLE 1. Category information and amount of each dataset.

Dataset	Category	No. training set	No. testing set
BDD100K [21]	*Bus; traffic light; traffic sign; person; bike; truck; motor; car; train; rider*	70,000	20,000
Pedestrian	*Crossing; not-crossing*	3,400	400
Vehicle	*Danger; no danger*	900	100
Traffic light	*Green light; green_left; green_up; green_right; red light; red_left; red_up; red_right*	5,240	400

Besides, there are some lightweight CNN models, such as MobileNet and EfficientNet. MobileNet is a small and efficient CNN model, which is suitable to install in real-time applications [15]. EfficientNet was first proposed in 2019 and optimized in terms of speed and accuracy [16].

IV. DATASET

This study mainly includes four tasks: object detection, dangerous vehicle prediction, intention recognition of pedestrian, and traffic light recognition. The four datasets towards different tasks are concluded as in table 1, and the following subsections introduce the used datasets in detail. The dataset called BDD100K [21] is available online, and the other datasets are collected and validated by ourselves. The size of all images in BDD100k dataset is 1280×720, and the size of images in three collected datasets is random. There are some sample images from the datasets used in this study, as shown in Fig. 6. Some sample images under different environmental conditions are shown in the first row. In the second row, the first three images are from the dataset with skeleton features, the last three images are from the dataset with only appearance features. The images in the third row are from the vehicle dataset. In the fourth row, there are eight types of the traffic lights, which are named 'green light', 'green_left', 'green_right', 'green_up', 'red light', 'red_left', 'red_right', 'red_up', respectively.

A. BDD100K DATASET

BDD100K includes 10 types of objects under different weather conditions or in different types of scenes, which is a large-scale and diverse dataset [21]. In this paper, the dataset containing 70,000 annotated images is used in the training process of the object detection part.

B. PEDESTRIAN DATASET

In the intention prediction task, there are 740 useful frames extracted from the videos that are captured by the driving recorder of a Tesla car. The extracted frames are from different sequences, so all of them are independent. By using the object detection algorithm, the areas of pedestrians are extracted from the collected images. After that, the PAFs algorithm was used to fit the skeletons for each pedestrian. The dataset used for the intention prediction consists of 1,900 images with fitted human skeletons and 1,900 images without fitted human skeletons.

C. VEHICLE DATASET

The various vehicles in the process of driving are recorded by the dash cam of a Tesla in the traffic scene. Firstly, the frames with vehicles are extracted from different sequences of the collected videos, and then the areas of vehicles are obtained and analyzed by using a deep learning-based method. The total number of images is 1,000, and the divided ratio between the training set and testing set is 9:1.

D. TRAFFIC LIGHT DATASET

In this paper, the dataset used for the traffic light task includes 8 kinds of common traffic lights. The dataset was manually collected from various websites and contains 5,240 training images and 400 testing images.

V. EXPERIMENTAL RESULTS

All experiments were worked on a Linux machine pre-installed with Ubuntu 14.04. It has four Titan X 12 GB GPUs, 64 GB of DDR4 RAM, and an Intel®Core i7-5930K processor. And the system is developed by using python programming language.

A. OBJECT DETECTION

As the main tasks in this study, both the precision and speed of object detection should be concerned. In this section, several standard detectors with state-of-the-art performance are evaluated from the aspect of detection accuracy (mAP) on the same dataset. On the other hand, the parameters of the optimized YOLOv4 are reduced by 74% by using the channel and layer pruning algorithm. Experiments indicate the inference time of the modified model (0.021s) is 36% (0.012s) faster than the original YOLOv4 (0.033s). The detection speed of the proposed system satisfies the real-time capability in terms of computational complexity.

Table 2 shows a comparison of detection accuracy (mAP) based on the BDD100K validation dataset. The models assessed include Single Shot MultiBox Detector (SSD) [20], Wasserstein Loss based Model for Object Detection (WLOD) [26], YOLOv3 [25], YOLOv4 [18], and the optimized YOLOv4. It is remarkable that the Mish function in the backbone of the YOLOv4 model is replaced with leaky-ReLU because the experimental environment used for this research cannot afford the large computation caused by the Mish function in the training process. The modified YOLOv4 outperformed other models because it not only used some effective data augmentation techniques in the data acquisition process, such as MixUp, Mosaic, and CutMix, but also adopted the intensive training method. By using these techniques, the training set was enriched to improve the performance of the model. The mAP of the modified YOLOv4 achieved the highest value of 52.7, which is 2.6 points higher than the original YOLOv4.

TABLE 2. Comparison of detection accuracy based on the BDD100K dataset, including YOLOv3, SSD, WLOD, YOLOv4 and Modified YOLOv4.

	YOLOv3	SSD	WLOD	YOLOv4	Modified YOLOv4
mAP	25.8	33.9	34.3	50.1	**52.7**

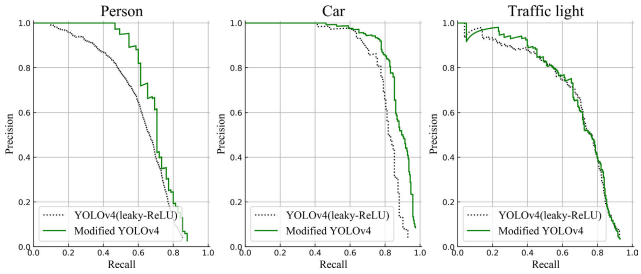


FIGURE 7. Precision-recall curves for three objects (person, car, and traffic light) on the BDD100k dataset. Compared YOLOv4 with the modified YOLOv4.

As a standard approach, the precision-recall (PR) curve is used to assess the performance of the experimental models. Person, cars, and traffic lights are the most common and important objects in the real traffic scene, which should be paid more attention than other objects. In this section, the detection performance of the modified YOLOv4 was compared with that of the original YOLOv4 with leaky-ReLU function based on BDD100K dataset. Fig. 7 presents the comparison on PR curves for three different objects. If the overlap value of prediction and GroundTruth is more than 0.5, then the sample is considered as true positive. Experiments shows that the average precision values of the modified model are significantly improved to 62.1% and 81.4% for person object and car object. For traffic light, the two experimental models achieved similar performance because the extracted images for intensive training do include a lot of traffic light samples.

B. SKELETON-BASED INTENTION RECOGNITION

For each detected pedestrian, pose estimation is applied to adjust a skeleton by using an efficient method named PAFs

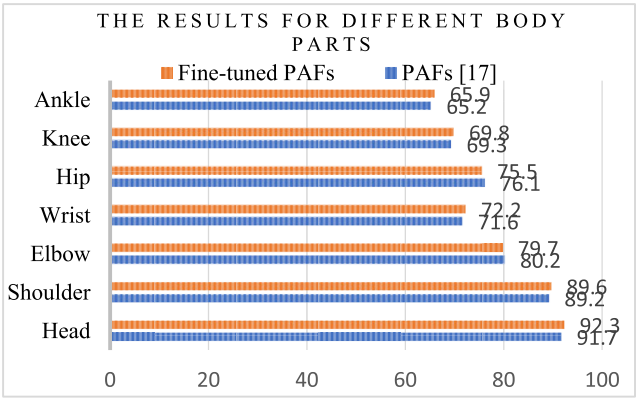


FIGURE 8. Comparison results between two experimental methods for seven different body parts.

with open source code [17]. In this study, the overall speed of the developed system is important to deal with multiple tasks. Thus, the original architecture is modified to improve the performance of the model in terms of the inference time. After conducting several experiments, the fine-tuned PAFs model uses SGD as the optimizer. The initial learning rate and momentum are set to 0.00005 and 0.9, respectively. The batch size is 10, and weight decay is 0.0001.

The original PAFs model and the fine-tuned model were evaluated on the MPII dataset [19], which is around 2000 samples excluded from the training set. Fig. 8 shows the assessment results of the experimental models, including the pose estimation of seven human parts (head, shoulder, elbow, wrist, hip, knee and ankle) and the inference time on the cropped image with a single person. Compared with the original PAFs method, the fine-tuned PAFs method enhanced the learning for some body parts that is difficult to locate. For example, the average precision of the ankle part obtained by fine-tuned PAFs is 0.7 higher than that of the original one. After modifying the PAFs method, precision of elbow and hip parts are lower than before among seven body parts. The inference speed of the original PAFS method under our test environment is much slower than the speed mentioned in the previous paper [17]. The average inference time of the fine-tuned PAFs method for an image with a single person is 0.09s, which is 30% (0.040s) faster than the original PAFs method (0.130s).

Intention recognition is realized by using a deep learning model on two distinct datasets (skeleton-based data & appearance-based data). Table 3 presents information on the datasets used in the intention recognition section and the performance of two CNN models for each dataset. The total number of images and the split ratio between training and testing are exactly same for the two datasets. Compared with the appearance-based data, both models achieved higher accuracy on the skeleton-based data. The main reason for this is that the skeleton features can provide useful information to the learning process. Moreover, the performance of a lightweight model (mobilenet) is compared with a model (VGG-19) that has a complicated structure in this experiment.

TABLE 3. Details of datasets and performances (recognition accuracy & time for each image) of two CNN models based on the datasets. "acc" refers to the recognition accuracy.

Pedestrian dataset	# of training set	# of testing set	VGG-19		MobileNet	
			acc	time	acc	time
With skeleton	1,700	200	97.5%	0.020s	90.5%	0.011s
Without skeleton	1,700	200	92.0%		83.1%	

The recognition time of mobilenet is 45% (0.009s) faster than that of VGG-19 for each image (0.020s). However, VGG-19 obtained the highest recognition accuracy of 97.5% based on the skeleton features, which is 7% higher than mobileNet on the same data. In the fine-tune process of VGG-19 model, the optimizer was set to Adam. The number of total epochs was 30, and the learning rate was 0.0001. Also, the first 17 layers of the model were frozen during the training process.

C. RISK ASSESSMENT

In the risk assessment stage, it mainly includes the dangerous vehicle assessment section and traffic light identification section, both of which are addressed as CNN-based classification tasks. For dangerous vehicle estimation, vehicles are classified into safe vehicles or dangerous vehicles. For traffic light recognition, traffic lights are recognized into the safe signal (green light) and dangerous signal (red light). In addition, the XAI technology is applied to explain the classification results in this study.

1) DANGEROUS VEHICLES ESTIMATION

In this study, the vehicles are divided into two classes (dangerous vehicles and safe vehicles) using a CNN model. If some cars stop or change direction, suddenly, they will threaten the safety of other vehicles. And the stationary vehicles or the vehicles that keep going straight is considered the vehicles without danger signals. In order to achieve high recognition accuracy, several famous models are compared and evaluated in this section. The best performance was recorded for each CNN model by adjusting various hyper-parameters. As shown in Fig. 9, each model result is provided with different colors, and the model parameters are converged after around twenty epochs. The performance of the fine-tuned Efficientnet model achieved the best validation accuracy at 94% among the examined models, which is 8% higher than the final recognition result of Resnet model. The learning rate was set to 0.001. The optimizer used for the fine-tuned Efficientnet model is Adadelta, which dynamically adapts over time and has minimal complexity beyond stochastic gradient descent (SGD) [10], [29].

2) TRAFFIC LIGHT RECOGNITION

Traffic light recognition is a vital task in the field of automatic driving, which directly affects traffic safety and order. The optimal hyperparameters are tuned to achieve the best

TABLE 4. Confusion matrix for traffic light recognition on the traffic light dataset (8 classes).

Class	Green light	Green _ left	Green _ right	Green _ up	Red light	Red _ left	Red _ right	Red _ up	Accuracy (%)
Green light	50	0	0	0	0	0	0	0	100
Green _ left	0	49	0	0	0	1	0	0	98
Green _ right	1	0	47	3	0	0	0	0	94
Green _ up	0	4	0	46	0	0	0	0	92
Red light	1	0	0	0	49	0	0	0	98
Red _ left	0	0	0	0	0	47	0	3	94
Red _ right	0	0	0	0	0	0	48	2	96
Red _ up	0	0	0	0	1	2	0	47	94
Average accuracy									95.75

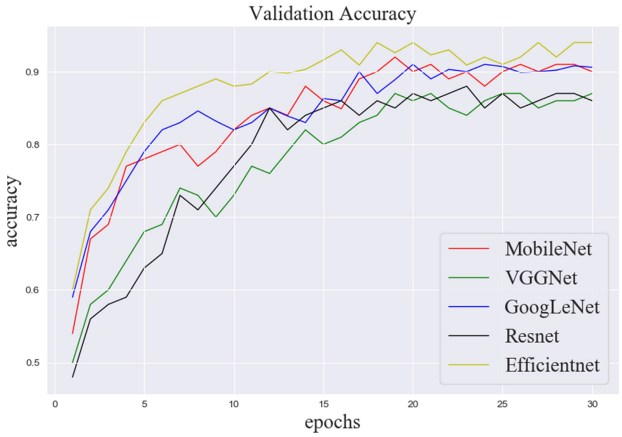


FIGURE 9. Plot of training epochs and validation accuracy for different CNN models models (MobileNet, VGGNet, GoogLeNet, Resnet and Efficientnet).

FIGURE 10. Saliency maps of testing images including different dangerous situations (brake, turn left, turn right, cross).

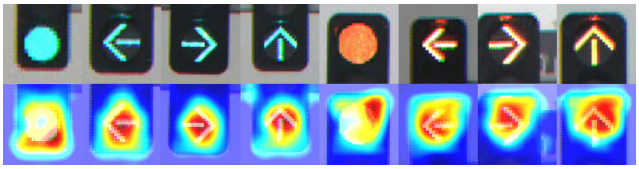


FIGURE 11. Saliency maps of testing images with 8 classes of common traffic light.

performance of MobileNet model in this experiment. For example, the learning rate was 0.001, the momentum was set to 0.9, and the optimizer was SGD. Table 4 presents in detail the confusion matrix results of the fine-tuned MobileNet model on the testing set with 8 classes. The experimental results suggested that model works well with an average accuracy of 95.75%. Since Green_ up and Green _ left have the same color and similar shape, the model misclassified the Green_ up as Green _ left, and the class named Green _ up has the highest error rate at 8%. Among 8 classes of common traffic lights, Green Light class obtained 100% recognition accuracy, and the other classes of traffic light have the accuracy between 94% and 98%.

3) EXPLAINABLE RECOGNITION RESULTS

In the risk estimation phase, CNN-based models are used to identify vehicles or traffic lights under current traffic situations. Neural networks learn features from the input image and give the corresponding classification results, which is considered a black-box algorithm. In this study, Randomized Input Sampling for Explanation (RISE) algorithm is applied to make the final classification explainable by generating the saliency map [11]. For dangerous vehicle estimation, the fine-tuned Efficientnet model is used to estimate whether the vehicles captured from the camera have dangerous behaviors. The dangerous behaviors are divided into four types, including brake, turn left, turn right, and cross. For traffic light recognition, the fine-tuned mobileNet recognizes 8 classes of traffic lights.

As shown in Fig. 10 and Fig. 11, the saliency map shows how important each pixel is for the final classification results. The light areas of the vehicle contribute to brake and orientation attention the most, while the tires parts are significant for the cross intention. Moreover, the saliency map highlights the importance of the circle parts to the red light and green light, as well as the importance of the arrow parts to each direction light.

D. QUALITATIVE EVALUATIONS

In this section, some qualitative results containing successful samples and failed samples are shown in Fig. 12 and Fig. 13, and then the performance of the whole system is evaluated from the aspects of accuracy and process time (See details in Table 5).

In this research, the testing images are captured by the Tesla driving recorder in the real traffic scene under different weather conditions [31]. From the visualized testing results

FIGURE 12. Visualized successful testing cases of the whole system in various traffic scenes under different environment conditions. 'C' means the pedestrian has the intention to cross the road. 'NC' means the pedestrian has no intention to cross the road. 'danger' indicates the vehicle has dangerous signals. 'normal' indicates the vehicle has no dangerous signals.

FIGURE 13. Visualized failed testing cases of the whole system in various traffic scenes under different environment conditions.

TABLE 5. Performance (accuracy and processing time) of four different tasks and the input size for each task.

Task	Performance		Time (s)	Input size
Object detection	mAP	52.7	0.021	608x608
Intention recognition (Skeleton-based)	accuracy (%)	97.50	0.020	224x224
Dangerous vehicles recognition		94.00	0.045	150x150
Traffic light recognition		95.75	0.010	32x32

of the proposed system, it can be observed the system can accurately detect most of the objects in each image, even some objects that are far from the camera. However, there are also some errors that occur in the testing process, as shown in Fig. 13. Although there are some images under different weather in the BDD100K dataset, severe weather still is a challenging factor in the testing process due to the low quality and bad illumination conditions. For example, in the first picture, the object detection model gives 3 boxes for the pedestrians, but there are 2 persons in the ground truth. The images with severe environmental conditions are considered to be handled to remove noise by some image preprocessing technique in the future, such as defogging, low-light enhancement. Since some of the pedestrians in the second picture are very small or hidden, it is hard to identify each body part of pedestrians in the phase of pose estimation, which directly affects the final results of intention recognition.

The performance of the entire system with multiple tasks are evaluated in terms of accuracy and the processing time of each task under the specific input size. As shown in Table 5, all recognition tasks can obtain good accuracy of over 94.00%. For the object detection task, the detection accuracy achieved 52.7% by using the optimized YOLOv4 model on BDD100K dataset. The processing time of the traffic light recognition task is the shortest among all tasks due to the lightweight MobileNet model and the small input size. According to the comprehensive evaluation of the system, all of the tasks can achieve a good performance in accuracy. However, since there are four different tasks in the developed framework, the total processing time per image is not fast enough for real-time applications.

VI. CONCLUSION

In this study, a vision-based object detection and recognition framework was proposed for autonomous driving. The proposed framework contains one object detection task and three recognition tasks. Various objects are detected by using an optimized YOLOv4 model with less parameters, which can achieve faster processing speed and higher detection

accuracy than the original. For detected objects, vehicles, pedestrians, and traffic lights are extremely important objects in the self-driving topic. Thus, there are three recognition tasks for the corresponding objects. By comparing with different CNN models, the most suitable model with the highest accuracy is selected for each recognition task. Besides, the RISE algorithm is used to explain the classification results by making the corresponding saliency maps for each image.

In the future, more attention should be paid to improving the overall speed of the proposed framework. To improve the performance of the system, a separate pipeline that can efficiently process single-frame-based and multi-frame-based recognition can be applied in the following study. Considering that the distance between the autonomous vehicles with other objects is important, distance prediction also should be integrated into this framework.

REFERENCES

[1] 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.

[2] S. H. Naghavi, C. Avaznia, and H. Talebi, "Integrated real-time object detection for self-driving vehicles," in *Proc. 10th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Nov. 2017, pp. 154–158.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[4] Y. Chen, D. Zhao, L. Lv, and Q. Zhang, "Multi-task learning for dangerous object detection in autonomous driving," *Inf. Sci.*, vol. 432, pp. 559–571, Mar. 2018.

[5] Z. Fang and A. M. Lopez, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Trans. Intell. Transp. Syst.*, early access, Oct. 18, 2019, doi: 10.1109/TITS.2019.2946642.

[6] M. Raza, Z. Chen, S. U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians' head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, Jan. 2018.

[7] R. Quintero Minguez, I. Parra Alonso, D. Fernandez-Llorca, and M. A. Sotelo, "Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1803–1814, May 2019.

[8] D. J. Phillips, T. A. Wheeler, and M. J. Kochenderfer, "Generalizable intention prediction of human drivers at intersections," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1665–1670.

[9] K.-H. Lee, T. Tagawa, J.-E.-M. Pan, A. Gaidon, and B. Douillard, "An attention-based recurrent convolutional network for vehicle taillight recognition," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 2365–2370.

[10] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*. [Online]. Available: http://arxiv.org/abs/1212.5701

[11] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*. [Online]. Available: http://arxiv.org/abs/1806.07421

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*. [Online]. Available: http://arxiv.org/abs/1905.11946

[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

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

[19] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[20] D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and C. Alexander Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[21] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2636–2645.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[23] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI*, 2020, pp. 12993–13000.

[24] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2736–2744.

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

[26] Y. Han, X. Liu, Z. Sheng, Y. Ren, X. Han, J. You, R. Liu, and Z. Luo, "Wasserstein loss-based deep object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, pp. 998–999.

[27] J.-G. Wang and L.-B. Zhou, "Traffic light recognition with high dynamic range imaging and deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1341–1352, Apr. 2019.

[28] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-CNNs for pose estimation and action detection," 2014, *arXiv:1406.5212*. [Online]. Available: http://arxiv.org/abs/1406.5212

[29] H. Wang, Y. Li, L. M. Dang, J. Ko, D. Han, and H. Moon, "Smartphone-based bulky waste classification using convolutional neural networks," *Multimedia Tools Appl.*, vol. 79, nos. 39–40, pp. 29411–29431, Oct. 2020.

[30] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107561.

[31] T. N. Nguyen, H. Nguyen-Xuan, and J. Lee, "A novel data-driven nonlinear solver for solid mechanics using time series forecasting," *Finite Elements Anal. Design*, vol. 171, Apr. 2020, Art. no. 103377.

YANFEN LI received the B.S. degree in software engineering from Linyi University, in 2018. She is currently pursuing the Ph.D. degree in computer science with Sejong University, Seoul, South Korea. She joined the Computer Vision Pattern Recognition Laboratory (CVPR Laboratory) in 2018. Her current research interests include computer vision, deep learning, image processing, and video coding.

HANXIANG WANG received the B.S. degree in software engineering from Linyi University, in 2018. He is currently pursuing the Ph.D. degree in computer science with Sejong University, Seoul, South Korea. He joined the Computer Vision Pattern Recognition Laboratory (CVPR Laboratory) in 2018. His current research interests include computer vision, video coding, and artificial intelligence.

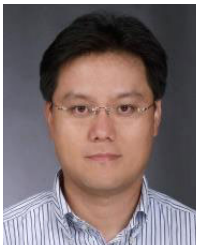
L. MINH DANG received the B.S. degree in information systems from the University of Information Technology, VNU HCMC, Vietnam, in 2016. He is currently pursuing the Ph.D. degree in computer science with Sejong University, Seoul, South Korea. He joined the Computer Vision Pattern Recognition Laboratory (CVPR Laboratory) at the beginning of 2017. His current research interests include computer vision, natural language processing, and artificial intelligence.

TAN N. NGUYEN received the M.E. degree from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, and the Ph.D. degree from Sejong University, South Korea, in 2019. He is currently an Assistant Professor with the Department of Architectural Engineering, Sejong University. His research interests include developing numerical methods, application robust methods to model structure considering modern materials, and new theoretical models. In addition, he also investigates highly nonlinear problems, instability of structures, and applies deep learning to structural analysis.

DONGIL HAN (Member, IEEE) received the B.S. degree in electronics and computer engineering from Korea University, Seoul, South Korea, in 1988, and the M.S. and Ph.D. degrees in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, Seoul, in 1990 and 1995, respectively. From 1995 to 2003, he was the Chief Research Engineer with Digital TV Research and Development Laboratories, LG Electronics Inc., Seoul. He is currently a Professor with the Department of Computer Science and Engineering, Sejong University, Seoul. His research interests include image processing, display quality enhancement for digital TV, system on chip, and robot vision.

AHYUN LEE, photograph and biography not available at the time of publication.

INSUNG JANG, photograph and biography not available at the time of publication.



HYEONJOON MOON received the B.S. degree in electronics and computer engineering from Korea University, in 1990, and the M.S. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo, in 1992 and 1999, respectively. From January 1996 to October 1999, he was a Senior Research with the Electro-Optics/Infrared Image Processing Branch, U.S. Army Research Laboratory (ARL), Adelphi, MD, USA. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) program. From November 1999 to February 2003, he was a Principal Research Scientist with Viisage Technology, Littleton, MA, USA. Since March 2004, he has been joined the Department of Computer Science and Engineering, Sejong University, where he is currently a Professor and the Chairman. His research interests include real-time facial recognition systems for access control, surveillance, and big database applications. He has extensive background on still image and real-time video based computer vision and pattern recognition. His current research interests include image processing, biometrics, artificial intelligence, and machine learning.