IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 1, 2022, accepted May 22, 2022, date of publication May 26, 2022, date of current version June 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178192

A Robust Layout-Independent License Plate Detection and Recognition Model Based on Attention Method

TAE-MOON SEO AND DONG-JOONG KANG
Department of Mechanical Engineering, Pusan National University, Pusan 43241, Republic of Korea

Corresponding author: Dong-Joong Kang (djkang@pusan.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant 2021R1A2C1010057.

ABSTRACT Auto License Plate Detection and Recognition (ALPDR) is required in many industrial fields, including visual surveillance systems and vehicle registration control. Even though this field has recently demonstrated high performance according to rapid advances in Deep Learning (DL) based technologies, there are still two problems that most related works have not yet solved. One is layout-dependent and the other is the problem of recognition accuracy degradation due to unconstrained environment conditions. In this paper, we present a more accurate, flexible and layout-independent method to improve LP detection and recognition accuracy under various outdoor conditions. For this, we proposed a new framework with lightweight and efficient anchor-free detection networks, employing the idea of CenterNet and attention-based recognition networks with residual deformable block suitable for LPR tasks. Various experiments demonstrate the performance of the proposed method, using famous LP benchmark datasets such as CCPD, AOLP and VBLPD. The proposed method outperformed the conventional LP detection and recognition methods. Additionally, Korean handicapped parking card (KHPC) data was tested to prove the usefulness of this method for marked character data different to license plates.

INDEX TERMS License plate detection, license plate recognition, attention, deep learning.

I. INTRODUCTION

Auto License Plate Detection and Recognition (ALPDR) is used in a wide range of fields, including vehicle management, digital surveillance systems, and intelligent transportations systems [1]. Recent advances in Scene Text Detection methods [2]–[4] have contributed to improvements in natural image analysis technology for character recognition. But in fact, there is a difference between those natural scene characters and marked characters, such as those used on LPs. For example, natural characters are comprised of various fonts, vocabulary, and rich semantic information. When training networks, these properties have a great influence on efforts to improve accuracy and performance. In contrast, marked characters, such as B-8-0, S-B-1, have little correlation between them. Therefore, a different approach is required than those used for natural scene characters. A conventional ALPDR system consists of three stages, (1) vehicle localization, (2) LP region detection, (3) character recognition. A typical method determines the LP region using object detecting neural networks such as Mask-RCNN [5], or Yolov3 [6], and then extract characters using character recognition networks such as CRNNs. Li *et al.* [7] proposed an end-to-end neural network that combined LP detection and recognition. Silva [8] proposed Warped Planar Object Detection Network (WPOD-Net) to detect LP images that were highly geometrically distorted. After localization of the vehicle using naïve Yolov2 [9], only the vehicle area is cropped from the image and resized to an image of normalized size based on the ratio of the LP's width to its height. The WPOD-Net detects the LP region and a modified OCR-network [8] performs character recognition. Xu *et al.* [10] offered large-scale Chinese License Plates. Roadside Parking Net (RPNet) is proposed to perform LP detection and recognition. However, despite these developments in recent decades, there are still challenges to applying them in real situations. A more advanced system is needed to cope with the various dynamic conditions of outdoor environments,

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

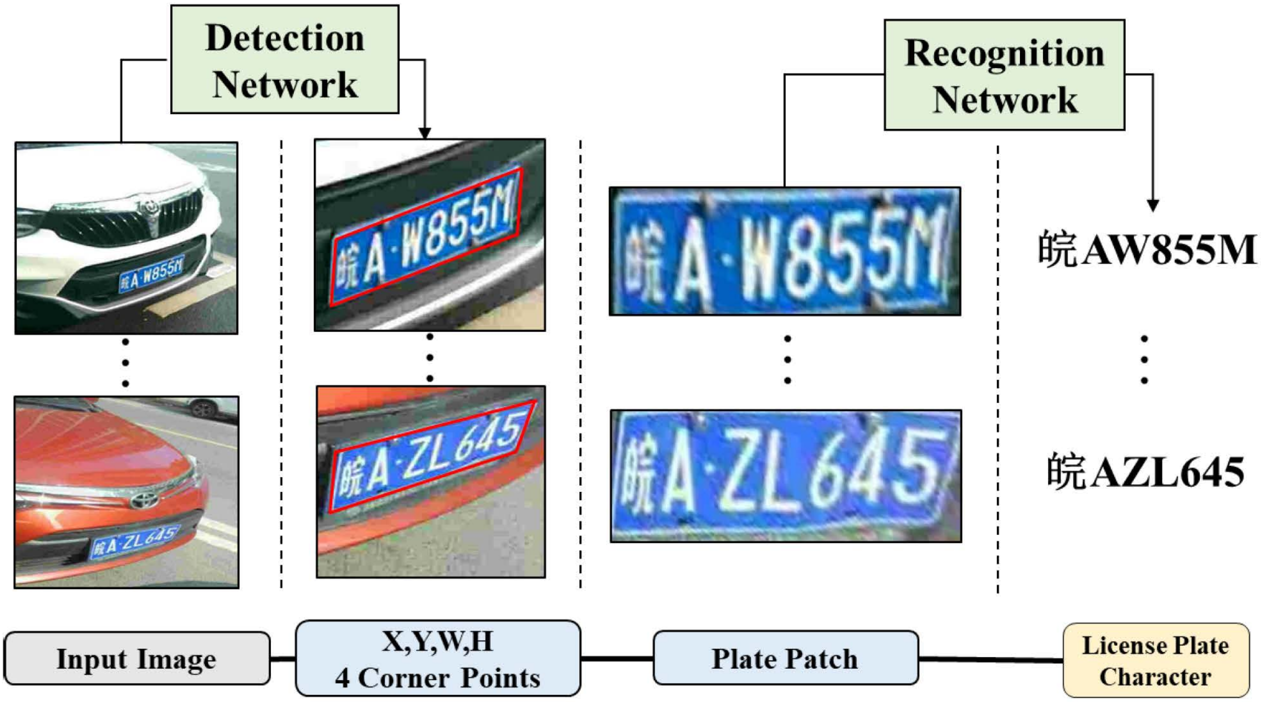


FIGURE 1. Overview of the proposed auto license plate detection and recognition system: The entire system consists of two modules from a detection network and a recognition network. Given the input RGB image, the detection module outputs LP bounding box and corner points coordinates. From this, a calibrated and normalized LP patch is created and the recognition network predicts LP characters.

including different resolution, background, location, lighting, and rotation changes. Also, the LP layout is different for each country or vehicle type. Therefore, in this paper, we propose a simple, efficient and flexible Anchor-free detection network inspired by the state-of-art general object detection method [11], [12]. We used this detection network to localize the LP bounding box and four corner points. Also, to perform noise-robust and layout-independent LP recognition, we design new feature extractor network with residual deformable block and redesign the recognition head [13] for irregular character recognition using an attention-based Transformer [14]. We achieved excellent performance with this LP recognition network. Our neural network was not limited to detecting and recognizing the LP. This makes it possible to use the proposed method for marked characters case other than those used for LPs. In our previous study [15] we identified the possibility of initial idea. We designed the license plate detection and recognition network architecture for experiment on CCPD dataset [10] and submitted the short paper to ICCAS'21 conference. Our work expands these previous studies.

The main contributions can be summarized as follows:

1. A flexible anchor-free LP character detection network was developed using a lightweight backbone, which efficiently locates the LP bounding box and four corner points.

2. We designed a new attention-based LP character recognition network with residual deformable block and first applied transformer to LP recognition, which can make possible layout-independent and robust-noise recognition.

3. By evaluating our method through experiments on various layout LP datasets and other mark character datasets, we demonstrated that our proposed model outperforms other methods and that our method is expandable for marked character data other than LPs

II. RELATED WORKS

The traditional approach to ALPDR is divided into two parts: LP detection (LPD) that finds the LP location, and LP recognition (LPR) that recognizes the characters on the LP. In this section, we divide LP detection and recognition into two subtasks and introduce them separately.

A. LICENSE PLATE DETECTION

The deep learning-based object detection methods usually extract features and regress location parameters using a deep neural network. These methods can be divided into three categories: (i) the anchor box-based two stage method, (ii) anchor box-based one stage method, and (iii) anchor-free method. Anchor boxes [16] are a set of predefined bounding boxes of a certain height and width in the position where the target object on the image may be located. A neural net predicts the bounding box by refining the anchor box using a regression method. In a two-stage method, representatively, such as the RCNN [17] series, a Region Proposal Network (RPN) [16] generates a Region of Interest (ROI). This is projected on a feature map and goes through a pooling step. A fully-connected head tunes the bounding box. The one-step method, such as a YOLO [18] series, is performed using a single neural network, by removing the steps of projecting and pooling the ROI. When an anchor box is used, it results in high detection performance and improves performance, especially for small objects. However, there are three drawbacks. First, the anchor box method needs a very large set of predefined bounding boxes. In Retinanet [19], more than 100K were used. From this, an imbalance problem arises between boxes belonging to the positive/negative categories. Second, in the process of creating an anchor box, we need to determine the parameters of a predefined box. This is a disadvantage in terms of an automation system that must minimize user intervention. Finally, anchor-based methods predict many overlapping bounding boxes, which must be reduced by Non-Maximum-Suppression (NMS). To address these shortcomings, a new object detection neural network that uses an anchor-free method [10], [12] has recently been proposed.

In this paper, using the idea of centernet, we adopt a general anchor-free object detection neural network for LP detection. This not only expands the LP applications, but can also be used for other marked character datasets with a sequential detection-recognition approach.

B. LICENSE PLATE RECOGNITION

It has been experimentally demonstrated that Convolution Neural Networks (CNNs) can learn the feature spaces of abundant data [20]–[22] and various optical character recognition neural networks based on CNN have been proposed [23]. Traditional methods regard LP recognition as a continuous label problem, and approach it using CTC loss (Connectionist Temporal Classification Loss) [24]. For LP recognition, a sliding-window type single class detector is also used [25]. In [26], focal CTC loss, which combines focal loss [19] with CTC loss, was used to solve the data imbalance problem for Chinese optical character recognition. The drawbacks of CTC loss are that the feature map of the network must be rearranged in consideration of real character sequences, and the performance is not good for noisy data that may occur in the actual data.

Crucially, in order to apply the mentioned methods to various layouts, a classifying module or network modification is needed. There are also commercial systems [27] that provide LP Recognition. However, users have difficulty to understand the details of system. Because only a part of the system is revealed, and good performance is only achieved with the target data. Because LP recognition requires high accuracy in real world, transfer learning [28] is essential for the target data set, so this unmodifiable black box type system has limitations.

In this paper, we modified and used the attention-based encoder-decoder method presented in [14] to design a noise-robust and layout-independent recognition network with residual deformable block.

TABLE 1. The architecture of the detection backbone network.

Input	Operator	Output
512 × 512 × 3	Conv 3x3 + BN + ReLU	512 × 512 × 64
512 × 512 × 64	Conv 3x3 + BN + ReLU	512 × 512 × 128
512 × 512 × 128	MaxPool	256 × 256 × 128
256 × 256 × 128	Residual Block	256 × 256 × 128
256 × 256 × 128	MaxPool	128 × 128 × 128
128 × 128 × 128	Residual Block	128 × 128 × 128
128 × 128 × 128	Conv 3x3 + BN + ReLU	128 × 128 × 256
128 × 128 × 256	Conv 1x1	128 × 128 × 1024

III. LICENSE PLATE DETECTION AND RECOGNITION NETWORK

A. FEATURE EXTRACTION FOR LICENSE PLATE DETECTION

Because our detection head needs to calculate the loss function in pixel levels, we designed a simple and light weight backbone network that does not collapse spatial information. We adopt some layers of ResNet18 [29] as feature extractors and used convolution operation, batch normalization [30], maximum pooling, dropout [31], and ReLU [32] as activation functions. The architecture of the feature extraction network is shown in Table 1. For LP detection, the features of the LP region are clear with respect to the input vehicle image. Thus, a single-scale simple network is sufficient to extract features. Given an input RGB image, it is resized to 512 × 512 resolution and passed through the backbone network. The size of the output feature map is down-sampled 4 times, and the channel is set to 1024, in order to richly express nonlinear features.

B. DETECTION HEAD

With inspiration from [11], we redesigned a detection head for the anchor-free method. As shown in Figure 2, the detection head takes the last stage of the feature extractor as an input and predicts outputs through two branches. For the input image I $\in R^{W\times H\times 3}$, the first branch predicts three outputs; a center point heatmap $\hat{Y}_{\text{heatmap}} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 1}$, box height and width $\widehat{Y}_{w,h} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 2}$, size offsets $\widehat{Y}_{\text{offsets}} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 2}$. Then the output is decoded to locate the bounding box. Also, the second branch predicts a three outputs corner point heatmap $\hat{Y}_{\text{corner}} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 4}$, corner point coordinates $\hat{Y}_{\text{coords}} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 8}$, coordinates offsets $\hat{Y}_{offsets} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 2}$. The second output is also decoded to locate the four corner

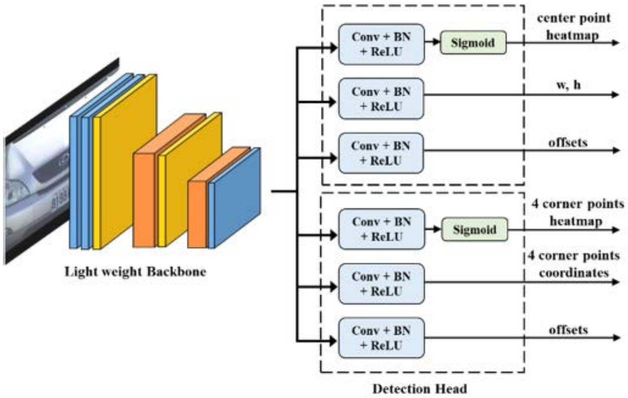


FIGURE 2. License plate detection module. Of the two output branches, the top is for the box and the other is for the corner points.

points. A rectified LP patch is obtained using the detected corner points.

C. FEATURE EXTRACTION FOR LICENSE PLATE RECOGNITION

The LP patch is rectified by using the LP position, which is the result of the detection network. The rectified image is $I_{\text{rectified}} \in R^{W\times H\times 3}$ and due to the characteristics of the LP, we set W = 100, H = 32 as a rectangular shape. Inspired by [33], we designed a new attention-based backbone network to focus on the character location in the feature map and to extract the features of the high-level characters. We suggest new backbone network that includes two Deformable Attention Stages with Residual Deformable Block. At this stage, as shown in Figure 3, the convolution filter is robust against spatial distortion as it has a reasonable receptive field by deformable convolution, and it can cover the multi-level feature through continuous downsampling and upsampling. In addition, by performing attention through the multiplication of low-level features and high-level features, it extracts character features efficiently. In conclusion, at this stage, because of attention based network that includes a Residual Deformable Block, our model can efficiently extract feature of irregular mark data such as multiple line, curved and unequal-sized words. The proposed network architecture is shown in Table 2.

TABLE 2. The feature extractor of recognition network.

Input	Operator	Output
32 × 100 × 3	Conv 3x3 + BN + ReLU	32 × 100 × 32
32 × 100 × 32	MaxPool	16 × 50 × 32
16 × 50 × 32	Residual Block	16 × 50 × 64
16 × 50 × 64	Deformable Attention Stage	16 × 50 × 64
16 × 50 × 64	Residual Block	8 × 25 × 128
8 × 25 × 128	Deformable Attention Stage	8 × 25 × 128
8 × 25 × 128	Residual Block	8 × 25 × 256
8 × 25 × 256	Residual Block	8 × 25 × 1024

D. RECOGNITION HEAD

Since the feature extractor doesn't include a fully connected layer, the created feature map retains spatial information. Here, c = 1024 is the channel of the feature map and k = $\frac{W}{4}\times\frac{H}{4}$ is the number of sequential feature tokens. Then, it passes through the recognition head to obtain the final decoded character. The recognition head is composed of three modules; a spatial relational attention module, a parallel attention module, and a character decode module. At this stages, the model calculates the similarity of position encodings through attention-based mechanism and predict output characters. This method enables to recognize layout-independent license plate as well as other marked characters. Spatial Relational Attention Module (Attn $_r$): In Eq (1)-(5), the feature map performs self-attention. This process makes a new feature map considering spatial similarity. Given a feature map $I \in \mathbb{R}^{k\times c}$ as the input of Attn$_r$, Position embedding vector (PE) of consecutive tokens Multi-Layer Perceptron (MLP), Multi-head Self-Attention (MSA), PositionWiseFeedFoward (PWFF), and Layer Normalization LN) methods are included. PWFF is another type of MLP. We set L to be 2. Output node is independent and can be optimized in parallel.

$$F_0 = I + PE \quad F, I, PE \in \mathbb{R}^{k\times C}. \tag{1}$$

$$F'_\ell = LN\left(MLP\left(MSA\left(F_{\ell-1}\right)\right)\right) \quad \ell = 1..L. \tag{2}$$

$$F_\ell = LN\left(PWFF\left(F'_\ell\right) + F'_\ell\right) \quad \ell = 1..L. \tag{3}$$

$$O = F_L \quad O \in \mathbb{R}^{k\times c}. \tag{4}$$

$$\therefore O = \text{Attn}_r(I). \tag{5}$$

The detailed process of MSA in Eq (2) is as follows [12]. SA is Standard **qkv** (query, key, value) self-attention. If X $\in \mathbb{R}^{k\times c}$ is an arbitrary sequential feature map, the learnable weight matrix $W_q, W_k, W_v \in \mathbb{R}^{c\times c}$, $W_o \in \mathbb{R}^{ch\times c}$, h is the number of heads, ch is $c \times h$.

$$X \cdot W_q = q, \quad X \cdot W_k = k, \; X \cdot W_v = v. \tag{6}$$

$$SA(X) = \text{Softmax}\left(\frac{q\cdot k^T}{\sqrt{\text{dim}(k)}}\right)\cdot v. \tag{7}$$

$$\text{MSA}(X) = \text{Concat}\left(SA_0(X), \cdots, SA_h(X)\right) W_o. \tag{8}$$

Parallel Attention Module (Attn $_p$): This module operation follows Eq (9). It multiplies the output of the spatial relational attention module with learnable weight, and then performs the attention using the output feature map of the feature extractor. In this process, both the features considering spatial relational attention and the previous features are used. Where

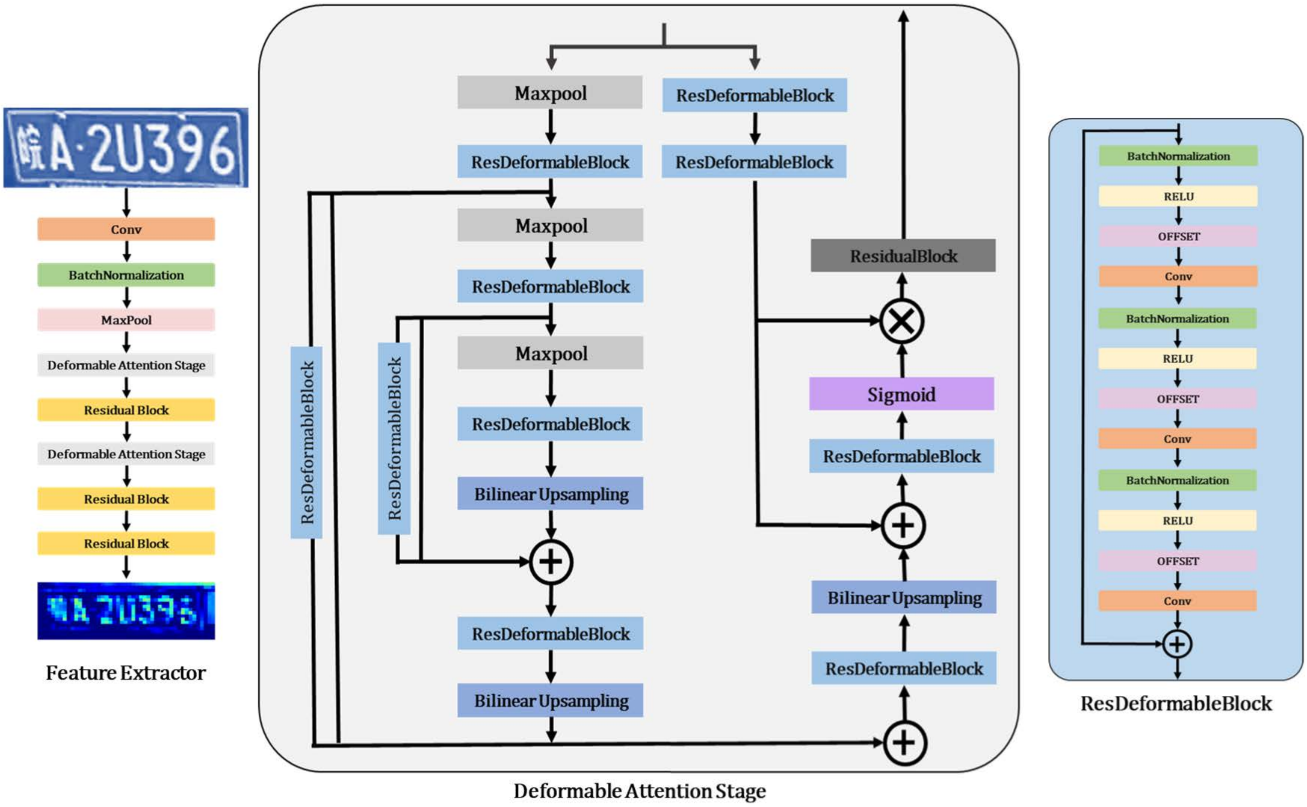


FIGURE 3. License plate recognition feature extractor.

$W_1 \in \mathbb{R}^{c\times c}$, $W_2 \in \mathbb{R}^{n\times c}$, n is the maximum recognizable character length.

$$G = \text{Softmax}\left(W_2 \cdot \tanh\left(W_1 \cdot O^T\right) \cdot I\right), \quad G \in \mathbb{R}^{n\times c} \tag{9}$$

Character Decode Module (CDM): The CDM decodes the previous results to the output character. Similar to the encoder, two layers of decoder layers are stacked for relational attention between output nodes. After that, CDM predicts the output characters through Softmax operation. This process is shown in Eq (10).

$$P = \text{Softmax}\left(W\left(\text{Attn}_r(G)\right)\right) \tag{10}$$

E. OPTIMIZATION

License Plate Detection: For the bounding box location, the network needs the position of the center, width, height and offsets of the LP for input image $I \in \mathbb{R}^{W\times H\times 3}$. The center point is given by heat map $Y_{xyc} \in [0,1]^{\frac{W}{4}\times\frac{H}{4}\times 1}$. This is created by applying a Gaussian kernel $Y_{xyc} = \exp\left(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right)$ to the coordinates of the object's center point $\tilde{p}_x, \tilde{p}_y$ when σ_p is the object size-adaptive standard deviation [10]. The center point loss function is a penalty-reduced pixel wise logistic regression with focal loss [19], as shown in Eq (11). It gives a small weight to a large number of easy negatives (background) and a large weight to a small number of difficult positives (keypoints). Therefore, it prevents the loss from being overwhelmed by a large number of negatives in the learning stage.

$$L_k = -\frac{1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right) & \text{if } Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha} & \text{otherwise}\\ \log\left(1-\hat{Y}_{xyc}\right) & \end{cases} \tag{11}$$

In this formula, N is the number of center points and we set 1, and α, β are hyperparameters for focal loss. In this paper, we use $\alpha = 2, \beta = 4$. Since the resolution of the input image is down-sampled by 4 ratios, spatial information is damaged. The offset term compensates for it. The offset loss is optimized through L1 Loss as shown in Eq (12).

$$L_{\text{offset}} = \frac{1}{N}\sum_{p}\left|\hat{Y}_{\text{offset}} - \left(\frac{p_{xy}}{4} - \tilde{p}_{xy}\right)\right|. \tag{12}$$

$$L_{\text{size}} = \frac{1}{N}\sum_{p}\left|\hat{Y}_{w,h} - S_k\right|. \tag{13}$$

Finally, the width and height of the bounding box is optimized. It is optimized using L1 Loss as shown in Eq (13). For offset loss and size loss, they were calculated only at the center point position. p_{xy} is the point position before down sampling. S_k is the width and height of the bounding box.

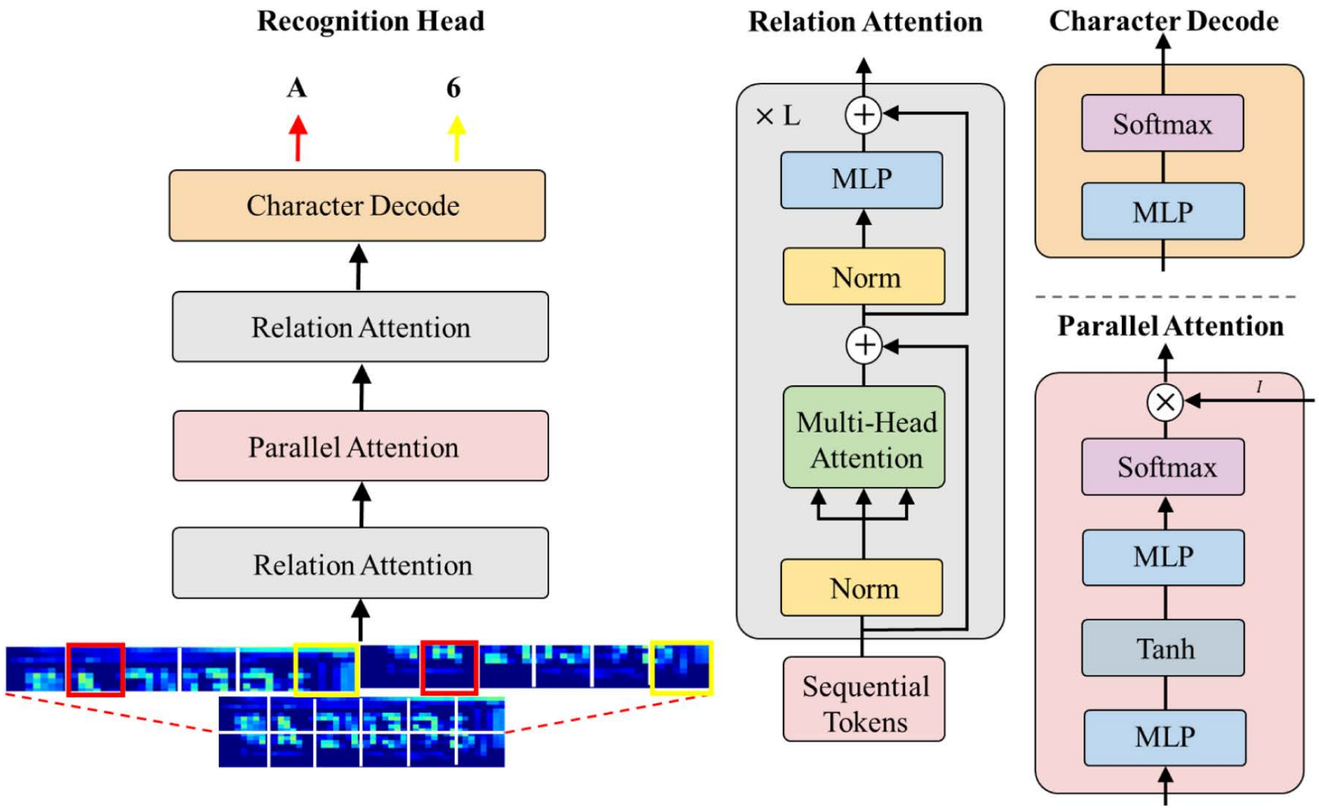


FIGURE 4. Spatial attention using the cross-attention mechanism of the transformer: It is composed of a spatial relation attention module, a parallel attention module, character decode module. The patches from the feature extractor are flattened and fed sequentially. For visual simplicity only 12 patches are expressed. Stack relation attention module L times. We use 2.

The function of the offset loss and the size loss are robust against the case of a large error by using L1 loss. Similarly, for the corner points, the loss is calculated by using Eq (11) for each of the corner points. L1 Loss is used to find the coordinates and offset of the corner point. Finally, the six loss functions are constructed by forming the joint loss function, considering the weights. We set the weight factor $\lambda_{\text{size}} = 1$, $\lambda_{off,b} = 0.05$, $\lambda_{off,c} = 0.05$, $\lambda_{\text{coord}} = 0.5$ in our experiments.

$$L_{\text{det}} = L_k + \lambda_{\text{size}} L_{\text{size}} + \lambda_{\text{o f f, b}} L_{\text{off, b}} + L_{corner} + \lambda_{\text{off, c}} L_{\text{off, c}} + \lambda_{\text{coord}} L_{\text{coord}} \tag{14}$$

License Plate Recognition: To recognize loss function, we adopted a cross-entropy loss at Eq (15). Where y_j denotes the true value and P_j means the predicted value.

$$L_{CE} = -\sum_{j=1}^{n} y_j \log\left(P_j\right). \tag{15}$$

IV. EXPERIMENTS

A. DATASETS AND SETTING

CCPD (Chinese City Parking Dataset) [10]: In order to improve detection and recognition performance, it is necessary to use a large dataset for training. Acquiring such data manually can be time consuming and costly. The recently released CCPD data is a total of 250k LP images. This dataset, collected from China, contains 7 sets of ccpd-base, ccpd-weather, ccpd-tilt ccpd-rotate, ccpd-fn, ccpd-db, and ccpd-challenge. Half of the ccpd-base were used for training, and the remaining 100k and other sub-datasets were tested.

AOLP (Application-Oriented License Plate) [34]: AOLP consists of 2049 images of Taiwan license plates. It is split into three sub-datasets: AC (access control, 681 images), LE (law enforcement, 757 images), and RP (road patrol, 611 images). Specifically, AC refers to cases where a vehicle passes a fixed passage at a reduced speed or with a full stop, LE refers to the cases where a vehicle is captured by a roadside camera, RP refers to the cases where a vehicle is captured by another moving vehicle. This dataset is used to demonstrate that the proposed network works robustly for a relatively small data set, not just a large data set. Half of the 2k images were used for training and the rest for testing.

VBLPD (Vietnamese Bike License Plate Dataset): In order to demonstrate the layout-independence of the proposed network, a two-line Vietnamese motorcycle car dataset was used. Since only the correct answer value for the location of the number region is provided, the labeling for the text was manually added. A total of 2k LP images obtained in a parking lot were used half for training and other for testing.

KHPC (Korea Handicap Parking Card): This is a dataset to verify the applicability of the proposed network to marked

TABLE 3. Description of sub-dataset in CCPD.

SubDataset	Quantity	Description
CCPD - base1(train)	100k	Images of cars in common scenes
CCPD - base2(test)	100k	Images of cars in common scenes
CCPD - weather	10k	Images taken on a rainy day, snow day or fog day
CCPD - tilt	30k	Images at horizontal tilt and vertical tilt
CCPD - rotate	30k	Images at horizontal rotate
CCPD - fn	20k	Images obtained from a relatively far or near
CCPD - db	10k	Images obtained in dark or extremely bright places
CCPD - challenge	50k	The most challenging image

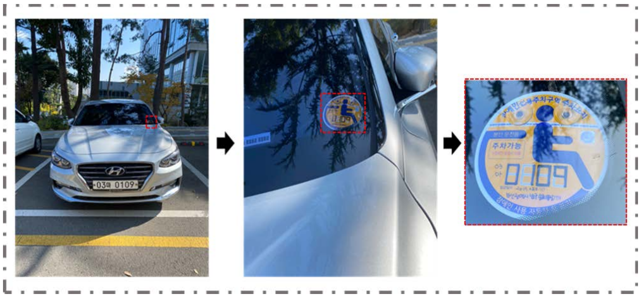


FIGURE 5. KHPC data acquisition process.

character problems in detection-recognition sequential methods other than LP. The handicap card is a parking card for the disabled person in Korea, and comes in white and yellow, with 4 numbers from 0 to 9 in the center area. Recognizing the 4 digits at center of the card is a problem. Due to the difficulties of collecting real data, 20k virtual images and 0.5k of collected real data were used to train through data synthesis, and only 0.5k of the collected real data was used for testing.

All of the experiments were performed using an Intel i-9 9900k CPU and NVIDIA QUADRO 8000 GPU. Our neural network was trained using the Adam [35] optimizer, and the running rate was initially set to 0.001 and then decreased by exponential strategy. In addition, a data augmentation technique was used to prevent overfitting. We applied transition, rotation ($-20° \sim 20°$), color jitter, blur for the data augmentation method. Later, we will release the source code.

B. EVALUATION CRITERION

To measure Detection Accuracy (DA), we used IoU [7] to calculate the value of the overlapping region of the real bounding box and bounding box predicted by our model. If the overlapping region is greater than or equal to the threshold, it is defined as True Detection (TD), and if it is not, it is defined as False Detection (FD). The performance of the detection accuracy was evaluated based on the threshold value $\lambda = 0.7$ as suggested in [10].

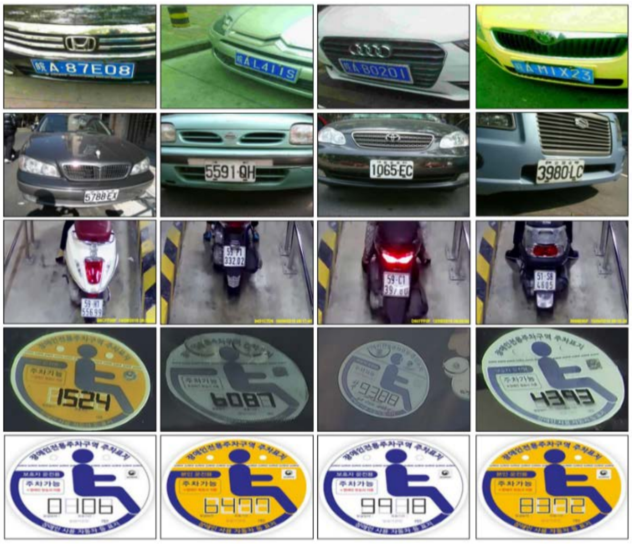
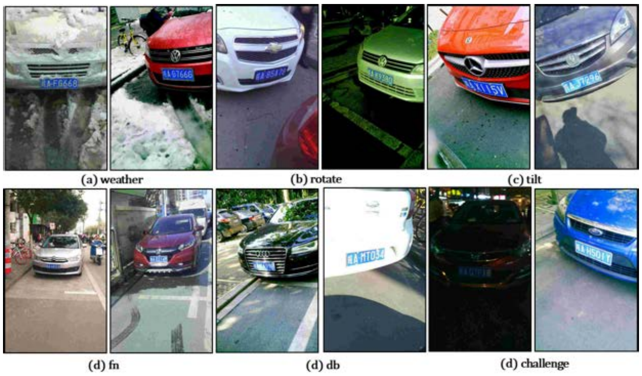


FIGURE 6. Examples of the annotated dataset: The first two lines at the top are the CCPD and AOLP datasets. And the next two lines are the VBPLD and KHPC datasets. The last line is the synthetic data for KHPC.



(a) weather (b) rotate (c) tilt (d) fn (d) db (d) challenge

FIGURE 7. Scenes where the proposed system successfully recognizes the license plate even in hard environmental conditions.

$$DA = \frac{TD}{TD + FD}(\%). \tag{16}$$

The Recognition Accuracy (RA) was evaluated only for the True Detection patch. Accuracy was calculated as in Eq (17). Matched character means the character was correctly predicted by the model.

$$RA = \frac{\text{Matched Characters \#}}{\text{Total Characters \#}}(\%). \tag{17}$$

V. RESULTS

A. RESULTS ON CCPD

The detection results and recognition results for the CCPD data are shown in Tables 4 and 5 respectively. Traditional experiments were also conducted using conventional

TABLE 4. License plate detection performance on the CCPD dataset.

Method	base (%)	weather (%)	tilt (%)	rotate (%)	fn (%)	db (%)	challenge (%)	Avg (%)
Edge-based [36]	91.64	91.53	90.29	90.29	90.51	90.38	89.68	90.62
MTCNN [37]	99.69	97.16	96.47	95.14	97.33	96.35	83.27	95.06
WPOD [7]	99.2	98.2	96.3	94.6	94.3	95.1	93.4	95.87
RPnet [9]	99.3	83.6	93.2	94.7	85.3	89.5	92.8	91.2
Ours	**99.94**	**99.49**	**99.2**	**99.20**	**98.02**	**99.27**	**94.8**	**98.56**

TABLE 5. License plate recognition performance on the CCPD dataset.

Method	base (%)	weather (%)	tilt (%)	rotate (%)	fn (%)	db (%)	challenge (%)	Avg (%)
Edge-based [36] + SVM	81.70	81.40	57.83	53.76	71.53	62.08	61.61	67.13
MTCNN [37] +LPRnet [38]	90.30	91.55	79.95	56.31	**90.11**	**86.89**	60.62	79.39
WPOD [7] + OCR [7]	90.76	90.88	**91.06**	92.21	64.88	82.86	64.40	82.43
RPnet [9]	92.36	89.53	87.83	86.51	65.16	84.43	62.25	81.15
SLPnet [36]	88.14	88.51	83.07	84.06	63.22	75.10	62.97	77.86
Ours	**99.83**	**97.48**	88.26	**94.11**	83.13	73.32	**75.38**	**87.36**

TABLE 6. Performance on the other datasets.

Method	AOLP (%)	VBLPD (%)	KHPC (%)
MTCNN [37] + LPRnet [38]	91.35	-	-
WPOD [7] + OCR [7]	**94.20**	-	-
RPnet [9]	91.85	-	-
Ours	92.30	**88.00**	**99.99**

methods, for comparison. Traditional edge detection algorithm [36] detect and cut the position of the LP. It uses the peak and valley of the histogram to segment the LP character and trains two SVM classifiers to recognize the target provinces and other characters on the LP. MTCNN [37] + LPRnet [38] is a light weight, open source ALPR framework. The LP region was detected using MTCNN [37]. Character recognition was performed using LPRnet. WPOD [8] + OCR [8] is a very novel neural network that recognizes performs recognition through OCRnet. RPnet [10] is a very good end-to-end detection and recognition neural network. SLPnet [39] is a recently proposed suffle block-based LPDR methodology. Plate detection is performed in an anchor-free manner in an end-to-end format, and then the region of interest is cropped to perform recognition.The above methods provided baseline performance for the CCPD data. Our network recorded 98.56% detection accuracy and 87.36% recognition accuracy. Our detection network demonstrated high performance with a small number of False-Positives by generating one predictive bounding box for one image. If sufficient data can be utilized, as in this experiment, using attention-based character inference can achieve high performance with noisy or distorted characters, unlike conventional methods for character inference using CTC loss. If transfer learning is used to learn by adding a small amount of sub dataset when learning, it is expected that higher performance will be possible.

B. RESULTS ON OTHER DATASETS

Table 6 shows the detection and recognition results for three types of data sets other than CCPD. These data were trained using only the bounding box information and text information, because the real corner points position was not known. While existing methods require modification of the model architecture or LP layout classification, the proposed network can be applied without modification by designing a general architecture. In addition, Since we use the Attention-based method, the recognition network can be directly applied to different layout LP data, and to mark characters other than the LP data without modification.

In order to verify this, we also experimented with AOLP, VBLPD and KHPC. The reason that AOLP data and VBLPD data show relatively low recognition performance compared to KHPC data is because there was an insufficient amount of data to train our recognition network. From several evaluations for different datasets, we prove that our proposed network can provide effectiveness for detection and recognition tasks on irregular mark shape data as well as conventional LP image. In real conditions, we can increase performance by training with enough data.

VI. CONCLUSION

In this work, we proposed a new framework consisting of an LP detection and an LP recognition network for layout-independent ALPDR. Figure 8 shows the results. We extracted only one license plate in one image for

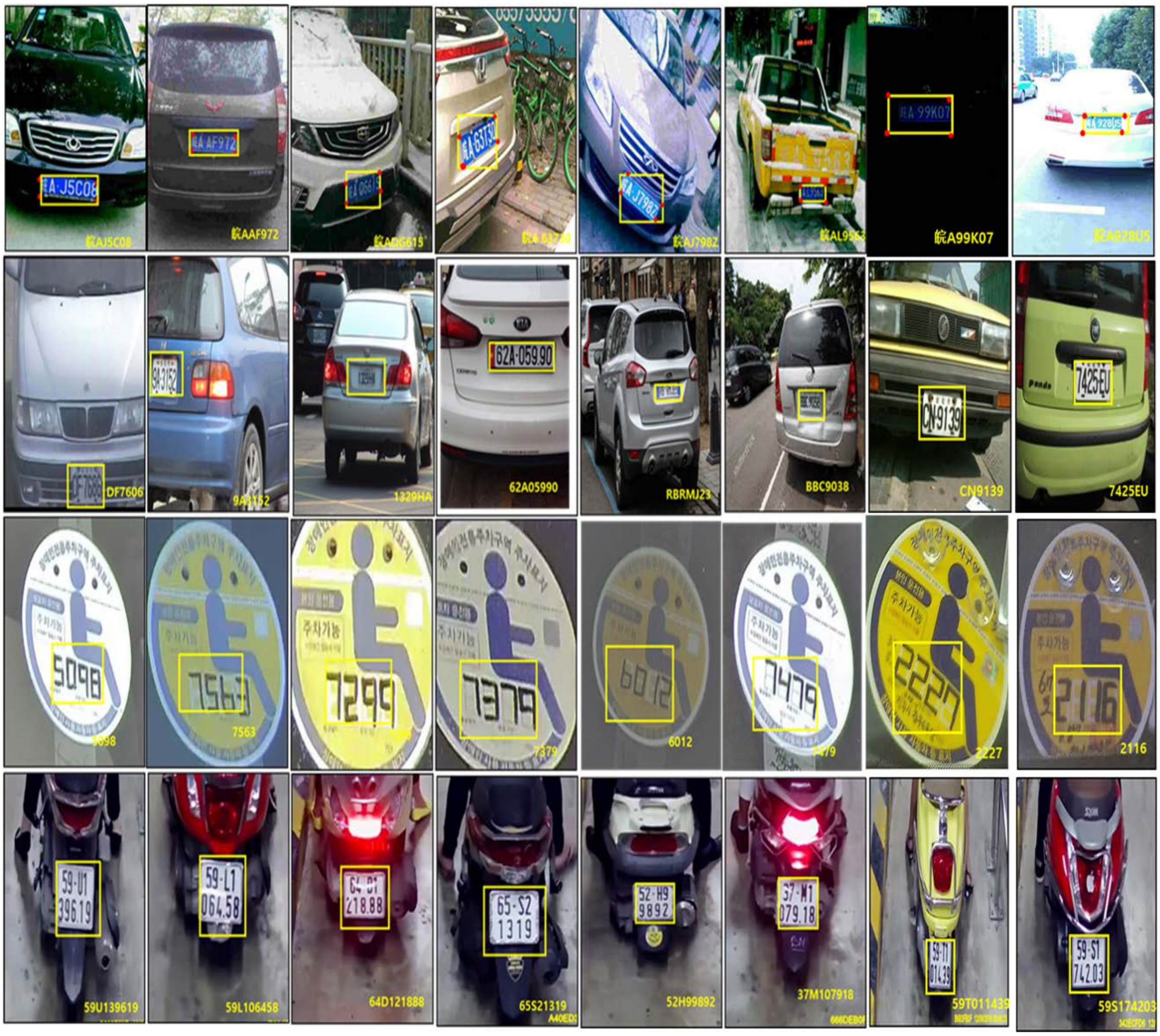


FIGURE 8. Detection and recognition results with our system.

comparison experiment. The LP detection network is composed of a simple and lightweight backbone network for feature extraction and a highly flexible heat map type detection head, through which it detects the bounding box and corner points of the LP. The recognition network outputs the result text using a patch obtained from the result of the detection network. The recognition network is composed of a feature extraction network including a Deformable Attention Stage and an attention-based recognition head, providing an operation that is layout-independent and robust to geometric distortion and noise. In addition, by designing a relatively general style detection and recognition network, it can be extended to marked characters other than LPs. Our framework is particularly robust against the spatial distortion and noise of LPs, and works regardless of layout. Tables 4 and 5 show that our system had higher detection and recognition performance for the CCPD dataset than conventional methods. In addition, Table 6 shows good performance for license plate data and marked character data with layouts different than those in CCPD. This means that it can be applied to a variety of marked character problems. In the future, we will be able to train our system in an end-to-end manner and expand it for use in more diverse scenarios such as videos [40] or applications [41].

REFERENCES

[1] B. Woodward and T. Kliestik, "Intelligent transportation applications, autonomous vehicle perception sensor data, and decision-making self-driving car control algorithms in smart sustainable urban mobility systems," *Contemp. Readings Law Social Justice*, vol. 13, no. 2, pp. 51–64, 2021.

[2] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9365–9374.

[3] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 71–79, 2018.

[4] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5551–5560.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[6] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[7] H. Li, P. Wang, and C. Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1126–1136, Mar. 2019.

[8] S. M. Silva and C. R. Jung, "License plate detection and recognition in unconstrained scenarios," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 580–596.

[9] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[10] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, and L. Huang, "Towards end-to-end license plate detection and recognition: A large dataset and baseline," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 255–271, 2018.

[11] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv:1904.07850*.

[12] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[13] P. Lyu, Z. Yang, X. Leng, X. Wu, R. Li, and X. Shen, "2D attentional irregular scene text recognizer," 2019, *arXiv:1906.05708*.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[15] T.-M. Seo and D.-J. Kang, "A new license plate detection and recognition model based on attention method," in *Proc. 21st Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2021, pp. 2191–2193.

[16] 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.*, vol. 28, 2015.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[18] 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.*, Jun. 2016, pp. 779–788.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[21] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 1133–1141.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[23] H. Li and C. Shen, "Reading car license plates using deep convolutional neural networks and LSTMs," 2016, *arXiv:1601.05610*.

[24] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[25] R.-C. Chen, "Automatic license plate recognition via sliding-window darknet-YOLO deep learning," *Image Vis. Comput.*, vol. 87, pp. 47–56, Jul. 2019.

[26] X. Feng, H. Yao, and S. Zhang, "Focal CTC loss for Chinese optical character recognition on unbalanced datasets," *Complexity*, vol. 2019, Jan. 2019, Art. no. 9345861.

[27] (2019). *Openalpr Cloud Api*. [Online]. Available: http://www.openalpr.com/2019

[28] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

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

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.

[34] G.-S. Hsu, J.-C. Chen, and Y.-Z. Chung, "Application-oriented license plate recognition," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 552–561, Feb. 2013.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[36] M. A. Lalimi, S. Ghofrani, and D. McLernon, "A vehicle license plate detection method using region and edge based methods," *Comput. Elect. Eng.*, vol. 39, no. 3, pp. 834–845, 2013.

[37] L. Xie, T. Ahmad, L. Jin, Y. Liu, and S. Zhang, "A new CNN-based method for multi-directional car license plate detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 507–517, Feb. 2018.

[38] S. Zherzdev and A. Gruzdev, "Lprnet: License plate recognition via deep neural networks," 2018, *arXiv:1806.10447*.

[39] W. Zhang, Y. Mao, and Y. Han, "SLPNet: Towards end-to-end car license plate detection and recognition using lightweight CNN," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Springer, 2020, pp. 290–302.

[40] N. R. Soora and P. S. Deshpande, "Color, scale, and rotation independent multiple license plates detection in videos and still images," *Math. Problems Eng.*, vol. 2016, Jul. 2016, Art. no. 9306282.

[41] G.-E. Abo-Samra, "Application independent localisation of vehicle plate number using multi-window-size binarisation and semi-hybrid genetic algorithm," *J. Eng.*, vol. 2018, no. 2, pp. 104–116, 2018.

TAE-MOON SEO received the B.S. degree from the School of Mechanical Engineering, Pusan National University, Busan, South Korea, in 2019, where he is currently pursuing the unified master's and Ph.D. degree. His research interests include deep learning, machine learning, and pattern recognition.

DONG-JOONG KANG received the B.S. degree in precision engineering from Pusan National University, Busan, South Korea, in 1988, and the M.S. and Ph.D. degrees in mechanical, and automation design engineering from the KAIST, South Korea, in 1990 and 1999, respectively. He is currently a Professor with the School of Mechanical Engineering, Pusan National University. His current research interests include machine vision, machine learning, and visual inspection in factory. From 2007 to 2019, he was an Associate Editor of the *International Journal of Control, Automation, and Systems*.