A Robust Attentional Framework for License Plate Recognition in the Wild

Linjiang Zhang, Peng Wang, *Member, IEEE*, Hui Li, Zhen Li, Chunhua Shen, *Member, IEEE*, and Yanning Zhang, *Senior Member, IEEE*

***Abstract*— Recognizing car license plates in natural scene images is an important yet still challenging task in realistic applications. Many existing approaches perform well for license plates collected under constrained conditions, *e.g.*, shooting in frontal and horizontal view-angles and under good lighting conditions. However, their performance drops significantly in an unconstrained environment that features rotation, distortion, occlusion, blurring, shading or extreme dark or bright conditions. In this work, we propose a robust framework for license plate recognition in the wild. It is composed of a tailored CycleGAN model for license plate image generation and an elaborate designed image-to-sequence network for plate recognition. On one hand, the CycleGAN based plate generation engine alleviates the exhausting human annotation work. Massive amount of training data can be obtained with a more balanced character distribution and various shooting conditions, which helps to boost the recognition accuracy to a large extent. On the other hand, the 2D attentional based license plate recognizer with an Xception-based CNN encoder is capable of recognizing license plates with different patterns under various scenarios accurately and robustly. Without using any heuristics rule or post-processing, our method achieves the state-of-the-art performance on four public datasets, which demonstrates the generality and robustness of our framework. Moreover, we released a new license plate dataset, named "CLPD", with 1200 images from all 31 provinces in mainland China. The dataset can be available from: https://github.com/wangpengnorman/CLPD_dataset.**

***Index Terms*— License plate recognition, attention mechanism, generative adversarial networks.**

I. Introduction

LICENSE Plate (LP) recognition in the wild is a fundamental problem in intelligent transportation systems. It can be used in a variety of applications including self-driving vehicles, traffic control and surveillance.

Manuscript received November 18, 2019; revised April 4, 2020; accepted May 13, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61876152 and in part by the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University under Grant CX2020184. The Associate Editor for this article was H. G. Jung. *(Linjiang Zhang and Peng Wang contribute equally to this work.) (Corresponding author: Peng Wang.)*

Linjiang Zhang, Peng Wang, and Yanning Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China, and also with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an 710072, China (e-mail: peng.wang@nwpu.edu.cn).

Hui Li and Chunhua Shen are with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: huili03855@gmail.com).

Zhen Li is with the Big Data and AI Technology Department, Minsheng Fintech Corporation Ltd., Beijing 100031, China.

Digital Object Identifier 10.1109/TITS.2020.3000072

Fig. 1. Examples of license plates successfully recognized by our proposed algorithm. (a) Dark Illumination; (b) Extremely bright or uneven; (c) Large horizontal tilt degree; (d) Large vertical tilt degree; (e) Images taken on a snowy or rainy day; (f) Mixture of bad conditions.

The LP numbers enable the link to a large body of information, including ownership, vehicle condition and driving record. Therefore, the technique of LP recognition in the wild can play a key role in road safety, traffic control and law enforcement. Although the recognition accuracy is acceptable for images shot under constrained conditions, recognizing license plates in complex environment is still far from satisfactory, especially for images photographed in dark, glare, occluded, rainy, snowy, tilted or blurred scenarios as shown in Figure 1.

With the advantage of deep neural networks, numerous work is proposed in recent years for license plate recognition, with Convolutional Neural Networks (CNNs) used for feature extraction, and Connectionist Temporal Classification (CTC) [1], number classifiers [2], *etc*. followed for character reading. These methods perform well for regular license plates (*eg*., nearly horizontal). When the license plate images are tilted or bent, an extra rectification step is required before recognition [3].

This paper tackles the task of license plate recognition in unconstrained scenarios. A robust framework is proposed to handle license plate recognition in both regular and challenging cases effectively. Our proposed license plate recognizer is composed of a 30-layer lightweight Xception for feature extraction and a 2D-attention based decoding module for character sequence recognition. Without extra processings like image rectification or character segmentation, the proposed model is capable of recognizing license plates in both regular and irregular patterns under various

practical scenarios. Different from current methods of treating a license plate as a one-dimensional sequence, our method uses 2D-attention that considers license plate image as a 2-dimensional signal. Trained in a weakly supervised manner, the proposed model is able to approximately localize the corresponding characters on license plates in decoding process, regardless of the appearance of license plate patterns.

Many license plate datasets are collected from one region, which causes bias in the datasets. For example, Xu *et al.* [2] introduce a license plate dataset **CCPD** which contains about 290K real world license plate images in various complex situations, as shown in Figure 1. However, since more than 95% of the images are photographed in one city, the first two characters in license plates are mostly the same, which may lead to bias for the trained model. In order to obtain a robust model which can be generally used for recognizing license plates from different regions, a CycleGAN model is tailored here which can mimic real scenarios and generate different kinds of license plate images, such as in dark or strong lighting conditions, containing shadows, *etc*. Moreover, license plates with various province characters can be synthesized, which alleviates the exhausting human annotation work to a large extent and enables a more general license plate recognition model. Our framework is evaluated on four public datasets. The competitive performance demonstrates the robustness of our framework. Moreover, we also collect a new license plate dataset with images from all 31 provinces in China, named "**CLPD**". It enables a more comprehensive evaluation of current plate recognition methods, and promotes the research of a more practical model.

It should be noted that the focus of this work is license plate recognition. So we simply train the off-the-shelf YOLOv2 detector [4] here to obtain bounding boxes of license plates.

The main contributions of this paper can be summarized as follows:

1. We design a robust method for license plate recognition in natural scene images. It is made up of a tailored Xception module and an encoder-decoder module. We optimized the recognition framework by using a 2D attention mechanism. It is able to extract local features for individual characters in a weakly supervised manner, without character level annotations needed. Compared to existing license plate recognition approaches, our method does not need an extra module to handle the irregularity of license plates or segment each character for recognition.

2. A tailored CycleGAN is proposed to synthesize license plates under various scenarios, including adding shadows, glare or darkness, perspective transformation, *etc*... With this engine we can generate license plate images with less data bias, and so get models with better generalization abilities.

3. We build a new dataset, named CLPD. It covers a large variety of photographing conditions, vehicle types and region codes, which provides a more comprehensive evaluation benchmark for plate recognition algorithms and promotes a more practical model design.

II. Related Work

In this section, we present a concise introduction to related works on license plate recognition, light-weight convolutional neural networks, generative adversarial networks and datasets of license plate.

A. License Plate Recognition

Existing methods for license plate recognition can be divided into two categories: Segmentation based [3], [5]–[8] and Non-segmentation based methods [1], [2], [9]. The segmentation based methods generally segment the license plate into characters and then recognize individual characters by OCR models [3], [5], [10]. Bulan *et al*. [10] perform segmentation and OCR jointly by using a hidden Markov models (HMMs) based probabilistic inference method, where the most likely character sequence is determined by Viterbi algorithm. Segmentation based methods rely heavily on the segmentation performance, which is very susceptible to the environment, including strong or weak lighting, bad weather, blurring, *etc*., and will result in a low recognition accuracy even with a strong recognizer.

Recent methods are mostly segmentation free. For example, Li and Shen [1] propose to treat license plate as a character sequence. Sequential features are encoded by CNNs and Bidirectional RNNs (BRNNs), and decoded by CTC without character separation. The CNN features are extracted from a well-trained CNN classifier, and the model cannot be trained end-to-end. RPnet proposed by Xu *et al*. [2] extracts ROI features from several different convolutional layers, and feeds the combined feathers to a series of classifiers for recognition. The number of classifiers is determined by the number of characters in the license plate, which limits its generalization ability in different regions. Li *et al*. [9] later propose a unified network which is able to localize license plates and recognize the letters at the same time in a single forward process. Similarly, the region features are encoded by BRNNs and decoded by CTC, which restricts its application to oriented LPs. Compared to the previous work, our method uses a 2D attention based encoder-decoder framework, where characters can be approximately localized by 2D attention regardless of LP image appearance, which enables its application to arbitrarily-oriented LPs.

B. Scene Text Recognition

License plate recognition can be regarded as a special case of general scene text recognition tasks, which have different characteristics. Characters in license plate usually use the same font in one region. There is no language model hidden in license plate, and no strong relationship with the context semantic information. In contrast, general scene text has a great variability on fonts. A language lexicon is existed and the text content is often highly relevant to the objects or scenes of the image. Xie *et al*. [11] propose a novel method where aggregation cross-entropy (ACE) is used for sequence recognition, replacing the generally used CTC loss owing to its inconvenience in processing 2D problems. A multi-object rectified attention network (MORAN) for scene text

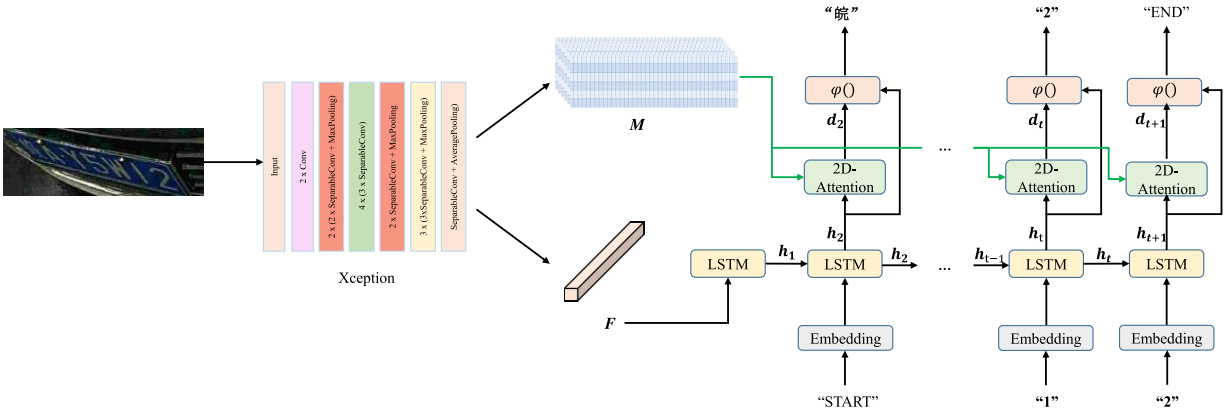


Fig. 2. Overview of the proposed architecture for LP recognition in complex scenarios. We extract license plates via a well-trained YOLOv2. The detected bounding box is fed into a 30-layer Xception network and get a global feature vector (denoted as F). An LSTM model is adopted to decode the obtained image feature into license plate numbers. We also extract an intermediate feature map (denoted as M) from the 12th layer of Xception, which provides local features during character decoding process.

recognition is proposed by Luo *et al*. [12], which contains a multi-object rectification network (MORN) and an attention-based sequence recognition network (ASRN). The image is rectified by MORN and then input to ASRN for recognition. Shi *et al*. [13] put forward a system that a flexible Thin-Plate Spline transformation is used to adaptively rectify a text image. A recognition model predicts a character sequence directly from the rectified image. Li *et al*. [14] use a 2D attention based encoder-decoder framework for irregular text recognition, which is very similar to our work. However, in our framework, a tailored CycleGAN is added for synthetic license plate generation, which can reduce data bias and improve model generalization ability.

C. Generative Adversarial Networks

With the invention of Generative Adversarial Networks (GANs) [15], many improved models have emerged, such as Deep Convolutional GANs (DCGANs) [16], Conditional GAN [17], Cycle-Consistent Adversarial Networks (CycleGAN) [18], Wasserstein GANs (WGAN) [19] *etc*. Zhu *et al*. [18] propose the CycleGAN, which learns the mapping between an input image and an output image using a training set of unaligned image pairs. In order to migrate the style of one image set to another one, cycle consistency loss is introduced. Based on this model, we propose an improved algorithm to generate synthetic license plate images in more complex environments, which improves the accuracy of license plate recognition furthermore. Wang *et al*. [20] adopt CycleWGAN to generate license plate images for improving recognition performance. Images simulating different shotting conditions are generated simultaneously. BRNN+CTC is used for plate recognition, which does not take oriented license plates into consideration as well. Nevertheless, we use a tailored CycleGAN to generate license plates under different conditions separately, which can lead to a better recognition performance.

D. Datasets of License Plate

Most datasets about license plates detection and recognition are collected from one area, and the type of license plate is monotonous (*eg*., only containing civic cars, no buses or trucks). Images are taken under similar conditions, such as highway toll stations and parking lots. Hence those datasets could not verify the robustness of a model.

Silva and Jung [3] collect a dataset named CD-HARD with 102 images, which covers some difficult situations, including tilting. However, because of the small number of images, the test result is susceptible to tricks. PKUData [21] captures images through a road surveillance camera, which includes a variety of license plate types and different lighting conditions. Unfortunately, all license plates are horizontal and taken from one province which has the same province code. Models trained on PKUData cannot be used to recognize license plates from other regions. AOLP [22] database consists of 2049 images with Taiwan license plate. This dataset is categorized into three subsets according to different levels of difficulty and photographing conditions. CCPD is currently the largest license plate dataset with 290k images, and is divided into multiple subsets such as tilt, difficulty, glare, and distance according to license plate conditions, which contributes greatly to the community. Nevertheless, more than 90% of the images are from one city too, which limits the trained model to recognize license plates from other areas. In this work, we propose to synthesize license plates by CycleGAN so as to make up for the deficiency. A new dataset names CLPD is introduced, which includes license plates from different provinces, to evaluate recognition models comprehensively.

III. Model

We introduce our proposed model in this section. As presented in Figure 2, the whole LP recognition model consists of two main parts: a tailored Xception network for feature extraction and a 2D-attention based RNN model for character decoding.

A. The Convolutional Image Encoder

A 30-layer Xception encoder is tailored from the original Xception [23] framework to fit our application, whose details are presented in Figure 3. The convolutional parts of our

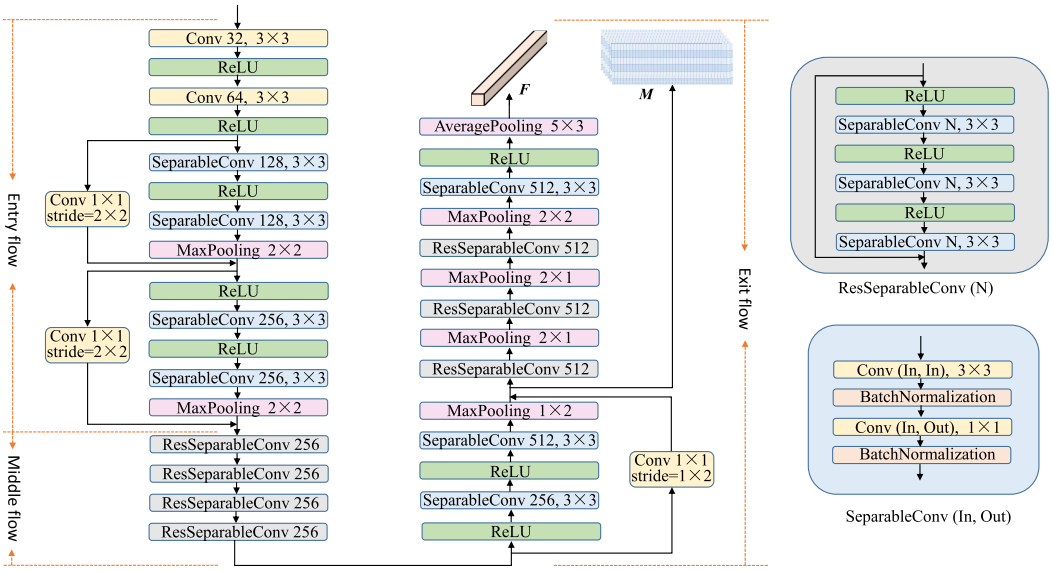


Fig. 3. The tailored Xception architecture. "Conv" stands for Convolutional layers, with output channels and kernel sizes presented. The stride and padding for convolutional layers are all set to 1, and no padding for Max-pooling layers.

model are based entirely on depthwise separable convolution layers [24]. The 30 convolutional layers are structured into 9 modules, where all of them have linear residual connections except for the first and the last one. The term "ResSeparableConv" stands for a stack of three separable convolution layers with an identity residual connection.

The entry flow downsamples the spatial size from 160×48 to 40×6 and increases the feature channel from 3 to 256 using interleaved separable convolutions and max-poolings. In the middle flow, we adopt repeated ResSeparableConv blocks to extract deep features that contain higher level representations, while the spatial size and channel number are fixed. In the exit flow, we extract a middle-level feature map M of size $40 \times 6 \times 512$ as context for attention network and a final feature vector F of 512 dimensions.

B. The Recurrent Sequence Decoder

RNN is widely used in translation, image caption, scene text recognition tasks. Here we extend it to license plate recognition. With a two-dimensional attention mechanism integrated, there is no need to make corrections for irregular license plate images or segment out each character for recognition. The proposed model can handle LPs in arbitrary shapes.

2-layer LSTMs with 512 hidden states each are adopted here in the sequence decoder. As shown in Figure 2, the holistic feature vector F is fed into LSTMs at time step 0, which aims to provide an overall information about the input image. Then a "START" token is input into the model at time step 1. From time step 2, the output of previous time step is fed into LSTMs until the "END" token received. The inputs of LSTMs are embedded by one-hot vectors followed by a linear transformation. The calculation of a single LSTM cell in training can be expressed as:

$$h_{t+1} = \mathrm{f}(h_t, \psi(x_t)), \qquad t = 1, \ldots, 8. \tag{1}$$

where h_t is the current hidden state, $\mathrm{f}()$ represents the LSTM operation at each time step and $\psi(\cdot)$ is the embedding operation. In inferring process, $x_t = \mathrm{softmax}(\varphi(h_t, d_t))$ which is the current output, while in training stage, the groundtruth character is adopted directly as x_t. $\varphi(\cdot)$ is a linear transformation, and d_t is the output of the 2D-attention module, which is calculated as follows:

$$\begin{cases} g_{ij} = \tanh(W_m M_{ij} + W_h h_t), \\ \alpha_{ij} = \mathrm{softmax}(W_g \cdot g_{ij}), \\ d_t = \sum_{i=1}^{H} \sum_{j=1}^{W} \alpha_{ij} M_{ij} \end{cases} \tag{2}$$

where M_{ij} is the feature vector at position i, j in M and h_t is the hidden state at time step t. W_m, W_h, W_g are linear transformation matrices to be learned; α_{ij} is the attention weight at location i, j; d_t is the weighted sum of image features, i.e., the local feature of the characters to be decoded at current time step t. The schematic of the 2D attention mechanism is illustrated in Figure 4.

IV. AsymCycleGAN for LP Image Generation

As aforementioned, it is difficult to manually collect LP images from a variety of regions, which makes most existing LP datasets heavily biased towards specific regional identifiers. In this section, we introduce a method for generating high-quality synthetic LP images using OpenCV and a tailored CycleGAN model (termed as AsymCycleGAN). With this approach, we are able to construct a balanced training data and reduce the reliance on manually collected data.

A. The Architecture of AsymCycleGAN

CycleGAN is an approach to translate an image from a source domain X to a target domain Y in the absence of paired training examples. In this work, the source domain X is composed of fake LP images generated by OpenCV and

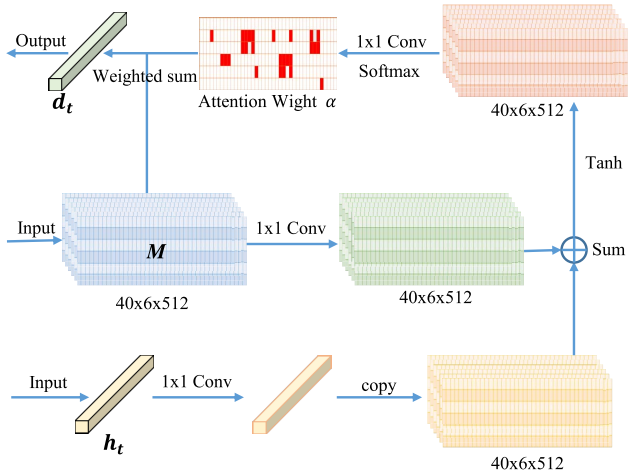


Fig. 4. The schematic of the 2D attention mechanism. M is the feature map of the image obtained by Xception (as shown in Figure 3), and h_t is the hidden state of each time step in decoding.

the target domain Y is made up of real LP images. There are four learnable modules in CycleGAN, leading to two mapping functions $G : X \rightarrow Y$, $F : Y \rightarrow X$ and two discriminators D_X and D_Y. The loss function of the standard CycleGAN can be expressed as follows:

$$\begin{aligned} L(G, F, D_X, D_Y) = & L_{GAN}(G, D_Y, X, Y) \\ & + L_{GAN}(F, D_X, Y, X) \\ & + \lambda L_{cyc}(G, F), \end{aligned} \tag{3}$$

where L_{GAN} represents the adversarial loss and L_{cyc} denotes the cycle-consistency loss:

$$\begin{aligned} L_{cyc}(G, F) = & E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] \\ & + E_{y \sim p_{data}(y)}[||G(F(y)) - y||_1]. \end{aligned} \tag{4}$$

In our case, what we need is the mapping function G to generate real images from synthetic images. $F(G(x))$ can be roughly regarded as generating a noisy image from a clean one and then remove these noises, while $G(F(y))$ is the opposite process. Note that in the process of $y \rightarrow F(y) \rightarrow G(F(y))$, the noise in y removed by F is in theory difficult to be exactly recovered by G, as one clean image can be associated to multiple real images with different noises. To this end, we replace the original cycle-consistency loss L_{cyc} (4) with

$$L_{cyc-new}(G, F) = E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1], \tag{5}$$

where the term with respect to $G(F(y))$ is removed. We term the modified CycleGAN model is AsymCycleGAN, as its cycle-consistency loss is asymmetric. The architecture of the proposed AsymCycleGAN model is shown in Figure 5.

B. AsymCycleGAN Generation Results

As in CycleGAN, the training of our proposed AsymCycleGAN model only requires two sets of unaligned images: synthetic and real images. As shown in Figure 6, the synthetic images are generated using OpenCV, while the real images are sampled from the CCPD dataset [2]. To generate different types of real images, we further divide the CCPD images

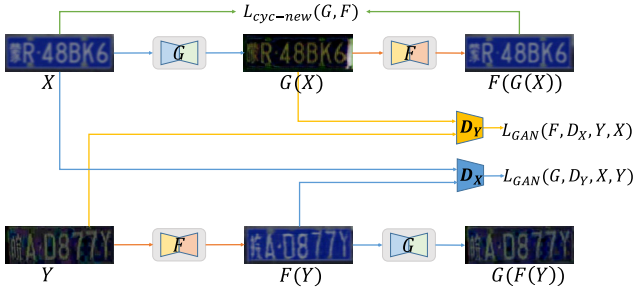


Fig. 5. The architecture of the proposed AsymCycleGAN model. X are synthetic LP images generated by OpenCV, Y are real LP images.

(a) Orignial LPs generated by OpenCV

(b) CycleGAN Images

(c) Asymmetrical CycleGAN Images

(d) Real Images

(e) Shadowed Image

Fig. 6. Various algorithms for generating license plate images. (a) Synthetic LPs generated by OpenCV; (b) The examples of LPs generated by CycleGAN model [18]; (c) The examples of LPs generated by our asymmetric CycleGAN model; (d) Real LPs from CCPD-DB; (e) Shadowed Image.

into two subsets with different illumination conditions: dark and bright. We use this dataset to train standard CycleGAN and our asymmetric CycleGAN model respectively, which consists of 800 synthetic LPs generated by OpenCV and 800 real-life license plate images in dark or glare environments. The AsymCycleGAN model is trained with a learning rate of 0.0002 and 30 epochs. The images generated by CycleGAN and asymmetric CycleGAN are shown in Figure 6. Moreover, we try to add shadows on the synthetic images so as to imitate real environment, the generated images are presented in Figure 6 (e).

V. The Proposed LP Dataset

In this chapter, we introduce a new LP dataset named CLPD (China License Plate Dataset), for a more comprehensive

Fig. 7. Sample images in our proposed CLPD dataset. Each license plate is manually annotated with a bounding box and its license number.

evaluation of LP detection and recognition algorithms, including how it is collected (Section V-A) and the comparison with other datasets (Section V-B).

A. Data Collection

The LP images in the proposed CLPD dataset are collected from a variety of real-scene image sources, for example, searched from the Internet, taken by mobile phones or captured by car driving recorders. All the faces shown in the images are blurred for privacy reasons. When taking LP photos, we also diversify the photographing angles, shooting times, resolutions and background so as to cover different conditions. The proposed dataset includes multiple vehicle types, such as trucks, cars, police cars and new energy vehicles. Note that new energy vehicles in China have license plates with eight letters, while other vehicles have seven-letter license plates. We also allow occluded license plates which have less than seven visible letters. The variation in the length of license plate letters increases the recognition difficulty as well, and makes the rule based recognition methods infeasible. The bounding boxes and license plate letters are annotated manually. In summary, the CLPD dataset contains 1200 LP images from all 31 provinces in mainland China. Some examples are shown in Figure 7. To our knowledge, our proposed LP dataset is the only one that covers all mainland China provinces with real shotted images.

B. Dataset Comparison

As presented in Table I, we compare our proposed dataset with other LP datasets in several aspects. Although the size of our dataset is small, it contains the most number of region codes. As we collect LP images from multiple sources, the image sizes are not fixed, in contrast to other datasets. Furthermore, AOLP, CCPD and our CLPD contain tilted images, while PKUData does not. Finally, our dataset contains LPs from different types of vehicles, including police car, new energy car and truck, which further increases the diversity of LP styles.

VI. Experiments

In this section, we conduct extensive experiments to compare our license plate recognition method with the state-of-the-art recognition methods. To demonstrate the effectiveness of the proposed model, plenty of experiments are performed on 4 different license plate datasets.

TABLE I

A Comparison of Available Datasets for LP Detection and Recognition. LP Size Is the Average Size of All License Plate Areas in a Dataset

	AOLP [22]	PKUData [21]	CCPD [2]	CLPD (ours)
Year	2012	2016	2018	2019
#Images	2049	2253	$290k$	1200
#Region Codes	0	23	29	31
LP size	72×28	156×39	253×100	149×48
Tilted	✓	✗	✓	✓
Var in vehicle type	✗	✓	✗	✓

A. Datasets

CCPD [2] is currently the largest publicly available License Plate (LP) dataset that provides over $290k$ unique Chinese LP images with detailed annotations. This dataset is separated into different groups according to the difficulty of identification, the illuminations on LP area, the distance from the license plate when photographing, the degree of horizontal tilt and vertical tilt, and the weather (rainy, snowy or fog). Each category includes 10k to 20k images. CCPD-base consists of approximately $200k$ images, where $100k$ are used for training and the other half is for test. The other subdatasets (CCPD-DB, CCPD-FN, CCPD-Rotate, CCPD-Weather, CCPD-Challenge) are also used for test.

AOLP [22] database consists of 2049 images of Taiwan license plate. This dataset is categorized into three subsets according to complexity levels and photographing conditions: Access Control (AC), Traffic Law Enforcement (LE) and Road Patrol (RP). Since we do not have any other images with Taiwan license plate, we use any two of these subsets for training and the remaining one for test, similar to previous practices [1], [9], [25].

PKUData [21] is released by Yuan *et al*., which provides images for license plate detection. The license plate labels are not annotated and we labeled the 2253 images in this dataset. 1352 images are randomly selected for training and the rest 901 are used for test.

CLPD is our proposed LP dataset, which contains 1200 images across all provinces in mainland China, with different vehicle types included. The images in the newly proposed CLPD dataset are all real and cover a large variety of photographing conditions, vehicle types and region codes. They are only used for test to verify the practicality of LP recognition models.

B. Implementation Details

In this work, we mainly focus on license plate recognition. In order to get the bounding boxes of license plates, a YOLOv2 [4] detector is trained on the training set of CCPD. We set the IOU threshold to 0.6, and achieve a detection performance of $precision = 99.4\%$ and $recall = 99.5\%$ on CCPD test sets. For fair comparison, we use the same evaluation criteria as that in [2]. An LP recognition result is correct if and only if the IoU between the detection and the ground truth is greater than 0.6 and all characters of the LP are correctly recognized (including the region code).

The recognition network is trained with cross-entropy loss and ADAM optimizer without any pre-training. In the training

TABLE II

RECOGNITION ACCURACY (%) WITH DIFFERENT CNN CHANNELS ON CCPD SUBSETS. 512 CHANNELS WILL BE ADOPTED IN FURTURE EXPERIMENTS

CNN Channels	Base	DB	FN	Rotate	Tilt	Weather	Challenge
128	99.7	98.6	98.7	96.1	97.7	98.1	86.3
256	99.7	98.9	99.0	97.1	98.3	98.2	87.9
512	**99.8**	**99.2**	**99.1**	**98.1**	**98.8**	**98.6**	**89.7**
1024	99.3	98.0	97.8	94.2	96.4	97.4	83.4

TABLE III

COMPARING THE RECOGNITION ACCURACY (%) WITH DIFFERENT LAYERS OF XCEPTION ON CCPD SUBSETS. WE CHOOSE THE OPTIMAL 30-LAYERS XCEPTION FOR FEATURE EXTRACTION

Layers	Base	DB	FN	Rotate	Tilt	Weather	Challenge
15	98.8	98.6	98.4	96.4	97.9	97.8	87.6
20	99.2	98.8	98.7	97.1	98.4	98.3	88.6
25	99.6	99.0	98.9	97.6	98.5	98.6	89.0
30	**99.8**	**99.2**	99.1	**98.1**	**98.8**	**98.6**	**89.7**
35	**99.8**	99.1	99.1	**98.1**	98.7	**98.6**	89.2
40	**99.8**	**99.2**	**99.2**	98.0	98.7	98.5	89.4

process, we adopt a batch size of 24 and a learning rate of $1e^{-3}$ initially. The learning rate is multiplied by 0.9 at every $12,000$ iterations until it reaches to $1e^{-5}$. The heights of input images in a batch are fixed, while the widths are calculated according to the aspect ratios of original images. All the experiments are conducted on an NVIDIA GTX1080Ti GPU with 11GB memory.

C. Ablation Studies

To analyze our proposed framework in detail, in this section, we evaluate it with different settings on CCPD dataset.

1) Effect of CNN Structures: In order to analyze the impact of CNN capacities, we first experiment with different number of CNN channels and layers. As shown in Table II, using more CNN channels indeed improves the license plate recognition accuracy, and the performance is saturated when the channel number reaches 512. Experimental results with different convolutional layers are demonstrated in Table III. The 30-layer Xception performs better than models with less layers, but the performance does not significantly improve when further increasing the depth. Hereinafter, we use the 30-layer Xception with 512 channels.

2) Effect of Inaccurate Bounding Box: Secondly, we test the recognition performance with detected and ground truth bounding boxes respectively, to demonstrate the robustness of our algorithm. Note that the detected bounding boxes may not encompass the license plates exactly as the groundtruth. This experiment is conducted to show the effect of bounding box variance on recognition performance. As shown in Table IV, the recognition accuracy only drops slightly by using detected bounding boxes (smaller than 0.2% for all cases except the "Challenge" one), which validates the robustness of our algorithm to inaccurate bounding boxes. One of the possible reasons is that the adopted 2D attention mechanism makes our algorithm not heavily depend on accurate bounding boxes: at each character decoding step, the adopted attention module will extract the most relevant local feature for each character in 2D space, instead of relying on heuristics rules for character separation.

TABLE IV

RECOGNITION ACCURACY (%) BY USING DIFFERENT BOUNDING BOXES ON SUB-DATASETS OF CCPD. THE EXPERIMENTAL RESULTS SHOW SMALL GAP WHEN USING INACCURATE BOUNDING BOXES, WHICH DEMONSTRATES THE ROBUSTNESS OF OUR ALGORITHM

Bounding Box	Base	DB	FN	Rotate	Tilt	Weather	Challenge
by Detection	99.8	99.2	99.1	98.1	98.8	98.6	89.7
Ground truth	**99.8**	**99.4**	**99.3**	**98.2**	**98.9**	**98.7**	**90.1**

TABLE V

THE RECOGNITION ACCURACY (%) ON CCPD-DB, WITH DIFFERENT NUMBER OF REAL IMAGES AND DIFFERENT GAN MODELS ADOPTED. USING SYNTHETIC DATA GENERATED BY OUR PROPOSED ASYM-CYCLEGAN OFFERS BETTER PERFORMANCE. THE SUPERIORITY IS EVEN OBVIOUS IF THERE ARE A SMALL NUMBER OF REAL IMAGES

Training Data	CCPD-DB	Improvement
Real (20k)	96.1	
Real (20k) + CycleGAN (20k)	96.3	0.2
Real (20k) + AsymCycleGAN (20k)	96.3	0.2
Real (20k) + CycleGAN (200k)	96.6	0.5
Real (20k) + AsymCycleGAN (200k)	**96.8**	**0.7**
Real (50k)	97.9	
Real (50k) + CycleGAN (50k)	98.0	0.1
Real (50k) + AsymCycleGAN (50k)	98.1	0.2
Real (50k) + CycleGAN (200k)	98.3	0.4
Real (50k) + AsymCycleGAN (200k)	**98.4**	**0.5**
Real (100k)	98.8	
Real (100k) + CycleGAN (100k)	98.8	0.0
Real (100k) + AsymCycleGAN (100k)	98.8	0.0
Real (100k) + CycleGAN (200k)	98.9	0.1
Real (100k) + AsymCycleGAN (200k)	**99.0**	**0.2**

3) Effect of Synthetic Data: The last ablation study is on the effectiveness of generated synthetic data. Here we also compare the performance by using different GAN models. We train our model with different numbers of real and synthetic images (20k, 50k and 100k), and then test the performance on CCPD-DB dataset. As shown in Table V, using the synthetic data generated by our proposed AsymCycleGAN offers better improvements than using that generated by the original CycleGAN, which demonstrates the superiority of our proposed AsymCycleGAN.

In addition, when comparing the improvements by using different number of real images, it can be found that the synthetic data plays a more important role when the real data size is smaller.

We can also see that the improvement is reduced when using smaller number of synthetic images. Note that the cost of generating synthetic images is very cheap: they do not need human annotation and the generation speed is fast (about 1K/min). So we can easily employ massive synthetic data for training, to improve the accuracy of LP recognition algorithms.

D. Experiments on Existing Benchmarks

1) Results on CCPD: It can be seen from Table VI that our algorithm outperforms other algorithms in terms of the overall AP and most of subsets, using the same real training data. The only exception is that the method of Luo *et al.* [12] is better than ours on the rotate and tilt subsets. The reason may be that Luo *et al.* [12] adopts an STN-based [31] technique

TABLE VI

LP RECOGNITION ACCURACY (%) ON EACH CCPD TEST SET (NUMBER OF IMAGES IN PARENTHESES). WE ACHIEVE THE HIGHEST RECOGNITION ACCURACY COMPARED WITH OTHER ALGORITHMS, ESPECIALLY IN THE DATASETS WITH ROTATTION AND CHALLENGING LICENSE PLATES

Model #Images	Overall	Base (100k)	DB (20k)	FN (20k)	Rotate (10k)	Tilt (10k)	Weather (10k)	Challenge (10k)	Test time ms
Ren *et al.* (2015) [26]	92.8	97.2	94.4	90.9	82.9	87.3	85.5	76.3	57.6
Liu *et al.* (2016) [27]	95.2	98.3	96.6	95.9	88.4	91.5	87.3	83.8	25.6
Joseph *et al.* (2016) [4]	93.7	98.1	96.0	88.2	84.5	88.5	87.0	80.5	23.8
Li *et al.* (2017) [9]	94.4	97.8	94.8	94.5	87.9	92.1	86.8	81.2	310
Zherzdev *et al.* (2018) [28]	93.0	97.8	92.2	91.9	79.4	85.8	92.0	69.8	17.8
Xu *et al.* (2018) [2]	95.5	98.5	96.9	94.3	90.8	92.5	87.9	85.1	**11.7**
Zhang *et al.* (2019) [29], [28]	93.0	99.1	96.3	97.3	95.1	96.4	97.1	83.2	153
Luo *et al.* (2019) [12]	98.3	99.5	98.1	98.6	98.1	98.6	97.6	86.5	18.2
Wang *et al.* (2020) [30]	96.6	98.9	96.1	96.4	91.9	93.7	95.4	83.1	19.3
Ours (Real Data Only)	98.5	99.6	98.8	98.8	96.4	97.6	98.5	88.9	24.9
Ours (Real + Synthetic data)	**98.9**	**99.8**	**99.2**	**99.1**	**98.1**	**98.8**	**98.6**	**89.7**	

TABLE VII

THE RECOGNITION ACCURACY (%) ON SUB-DATASETS OF AOLP. OUR APPROACH PERFORMS BETTER THAN OTHER METHODS ON ALL THREE SUBSETS

Model #Images	AC (681)	LE (757)	RP (611)
Li *et al.* (2016) [1]	94.9	94.2	88.4
Li *et al.* (2017) [9]	95.3	96.6	83.7
Wu *et al.* (2018) [25]	96.6	97.8	91.0
Ours	**97.3**	**98.3**	**91.9**

TABLE VIII

THE RECOGNITION ACCURACY (ACC, %) AND RECOGNITION ACCURACY WITHOUT REGION CODE (ACC W/O RC, %) ON PKUDATA AND CLPD. FOR ACC W/O RC, THE RECOGNITION IS CONSIDERED TO BE CORRECT IF ALL THE CHARACTERS EXCEPT THE FIRST ONE REGION CODE ARE CORRECTLY RECOGNIZED

Dataset	PKUData		CLPD	
Criterion	ACC	ACC w/o RC	ACC	ACC w/oRC
Masood *et al.* (2017) [32]	-	89.3	-	85.2
Xu *et al.* (2018) [2]	77.6	78.4	66.5	78.9
Ours (Real Data Only)	84.8	86.5	70.8	86.1
Ours (Real + Synthetic Data)	**88.2**	**90.5**	**76.8**	**87.6**

which is specifically designed for rotated images. Note that our algorithm can also benefit from using this technique and the accuracy on the rotate and tile subset is expect to be further improved.

Our algorithm shows significant superiority on subsets with irregular LP images, such as "Rotate", "Weather" and "Challenge", which again proves the robustness of our model to the deformation of license plates. Moreover, by adding synthetic images generated by our AsymCycleGAN, the recognition accuracies consistently raise furthermore on all subsets (a 0.4% gain of Overall). The increment is even obvious when LPs are rotated or tilted (raising 1.7% on CCPD-Rotate and 1.2% on CCPD-Tilt). The main reason is that random perspective transformation and rotation are applied to the synthesized data, which is a great complementary to real data.

We select some extremely distorted images and visualize the 2D attention heat maps when decoding each character in Figure 8. The results show that even for very tilted images, the 2D attention model can locate to the character being decoded and extract corresponding features for recognition. It should be noted that the attention module does not require additional character-level annotations. It is trained in a weakly supervised manner by the cross-entropy loss on the whole plate recognition.

2) Results on AOLP: In this section, we compare our model with other state-of-the-art methods on AOLP dataset. For fair comparison, we did not use any synthetic data during model training. Perspective transformation is employed for data augmentation. The results in Table VII show that our approach performs better than other methods on all three subsets, which validates the superiority of our approach. In particular, our method leads to the accuracy increments of 0.7% on AC, 0.5% on LE and 0.9% on RP, compared to the second best results. Note that the RP subset is mainly composed of oriented or distorted license plates, on which our method obtains the largest performance gain. This result further demonstrates the effectiveness of our model in recognizing irregular license plates.

3) Results on PKUData: For PKUData, we randomly sample three-fifths for training and use the remaining two-fifths for test. For fair comparison, we re-train the model proposed in [2] by the same training data. An open API called Sighthounds [32] is tested as well, but we have no idea about the training data it used. We evaluate the LP recognition accuracy on two settings, *ie*., with and without region code (a Chinese character) considered. The model in Sighthounds [32] does not support region code recognition, so we only report its accuracy without region code. The recognition results are shown in Table VIII. Our model outperforms that in [2] by about 7% when only real data is adopted, and surpasses Sighthounds [32] if synthetic training data is added. In comparison with the improvement on CCPD dataset, the accuracy gain is even more obvious when using synthetic data (about 4%), because of the limited real training images in PKUData, which demonstrates the usefulness of our synthesis engine when there is scarce training data.

E. Experiments on Our CLPD Dataset

As aforementioned, the diversity of our proposed CLPD dataset is much larger than existing LP datasets, which provides a platform to evaluate current algorithms comprehensively. We train the proposed model on CCPD-Base dataset, and test it on CLPD. Experimental results in Table VIII show the advantage of our model. It leads to the highest accuracy no matter region code is considered or not. By adding synthetic

Fig. 8. Visualization of 2D attention weights at each decoding timestep. Results indicate that the 2D-attention model can handle challenging cases.

Fig. 9. Detection and recognition results on CLPD using YOLOv2 and our recognition model. With the addition of synthetic data, the model is able to recognize license plates from different provinces under various scenarios.

Fig. 10. Examples of LPs that are incorrectly recognized by the proposed method. The ground truth is shown in the parenthese.

data, the accuracy increases 6% further if region code is considered, which benefits from a more balanced region code distribution in our synthetic data that can be easily obtained by the proposed engine. Some experimental results are visualized in Figure 9.

We also present some failure cases in Figure 10. As there is no specific language rule used in license plate, some similar characters are rather difficult to be distinguished, such as "4" and "A", "8" and "B","0" "D", and "O". Images with extreme blur or occlusion are also unable to be recognized.

VII. Conclusion

In this paper, we present a robust model for license plate recognition in unconstrained environment. The proposed model is built upon an Xception CNN module for feature extraction, and a 2D-attention based RNN module for sequence decoding. To handle the shortage or unbalance of real training data, CycleGAN is tailored to generate synthetic LP images with different deformation styles and a more balanced region codes, which provides a simple yet effective way to complement available real data. Extensive experimental results indicate the superiority of our methods, especially when addressing distorted license plates or with limited training data. An LP dataset that contains images captured in different ways from various regions is collected so as to evaluate LP recognition methods more comprehensively.

We use an LSTM-based sequence decoder for license plate recognition, which cannot be trained in parallel over time steps. For future works, a transformer-like decoder may be explored to accelerate training speed.

References

[1] H. Li and C. Shen, "Reading car license plates using deep convolutional neural networks and LSTMs," 2016, *arXiv:1601.05610*. [Online]. Available: http://arxiv.org/abs/1601.05610

[2] Z. Xu *et al.*, "Towards end-to-end license plate detection and recognition: A large dataset and baseline," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 255–271.

[3] S. M. Silva and C. R. Jung, "License plate detection and recognition in unconstrained scenarios," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 593–609.

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

[5] C. Gou, K. Wang, Y. Yao, and Z. Li, "Vehicle license plate recognition based on extremal regions and restricted Boltzmann machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1096–1107, Apr. 2016.

[6] J.-M. Guo and Y.-F. Liu, "License plate localization and character segmentation with feedback self-learning and hybrid binarization techniques," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1417–1424, May 2008.

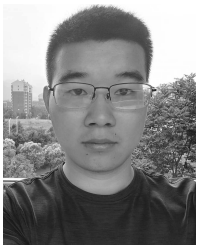
[7] G. R. Gonçalves, S. P. G. da Silva, D. Menotti, and W. R. Schwartz, "Benchmark for license plate character segmentation," *J. Electron. Imag.*, vol. 25, no. 5, Oct. 2016, Art. no. 053034.

[8] P. Li, M. Nguyen, and W. Q. Yan, "Rotation correction for license plate recognition," in *Proc. 4th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2018, pp. 400–404.

[9] 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.

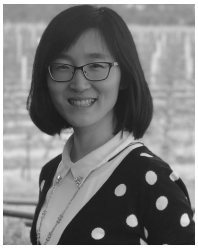
[10] O. Bulan, V. Kozitsky, P. Ramesh, and M. Shreve, "Segmentation- and annotation-free license plate recognition with deep localization and failure identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2351–2363, Sep. 2017.

[11] Z. Xie, Y. Huang, Y. Zhu, L. Jin, Y. Liu, and L. Xie, "Aggregation cross-entropy for sequence recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6538–6547.
[12] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.
[13] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.
[14] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8610–8617.
[15] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784
[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875
[20] X. Wang, Z. Man, M. You, and C. Shen, "Adversarial generation of training examples: Applications to moving vehicle license plate recognition," 2017, *arXiv:1707.03124*. [Online]. Available: http://arxiv.org/abs/1707.03124
[21] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, and N. Komodakis, "A robust and efficient approach to license plate detection," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1102–1114, Mar. 2017.
[22] 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.
[23] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[24] L. Sifre, "Rigid-motion scattering for image classification," Ph.D. dissertation, Center Appl. Math., École PolyTechn., Palaiseau, France, 2014.
[25] C. Wu, S. Xu, G. Song, and S. Zhang, "How many labeled license plates are needed?" in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Cham, Switzerland: Springer, 2018, pp. 334–346.
[26] 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.
[27] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 21–37.
[28] S. Zherzdev and A. Gruzdev, "LPRNet: License plate recognition via deep neural networks," 2018, *arXiv:1806.10447*. [Online]. Available: http://arxiv.org/abs/1806.10447
[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[30] T. Wang *et al.*, "Decoupled attention network for text recognition," 2019, *arXiv:1912.10205*. [Online]. Available: http://arxiv.org/abs/1912.10205
[31] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[32] S. Zain Masood, G. Shu, A. Dehghan, and E. G. Ortiz, "License plate detection and recognition using deeply learned convolutional neural networks," 2017, *arXiv:1703.07330*. [Online]. Available: http://arxiv.org/abs/1703.07330

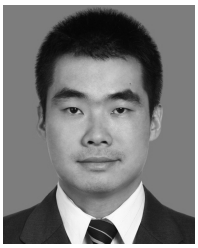


Linjiang Zhang received the B.E. degree in computer science and technology from Northwestern Polytechnical University, China, in 2017, where he is currently pursuing the M.E. degree with the School of Computer Science.

Peng Wang (Member, IEEE) received the Ph.D. degree in control science and engineering from Beihang University, China, in 2011. He was with The University of Adelaide for about four years. He is currently a Professor with the School of Computer Science, Northwestern Polytechnical University, China. His research interests include computer vision, machine learning, and artificial intelligence.



Hui Li received the Ph.D. degree from The University of Adelaide in 2018. She is currently a Research Fellow with the School of Computer Science, The University of Adelaide, Australia. Her research interests include scene text detection and recognition and visual question answering.



Zhen Li received the Ph.D. degree in pattern recognition and intelligent systems from Beihang University in 2014. He is currently the Director of the Big Data and AI Technology Department, Minsheng Fintech Corporation Ltd. His research interests include big data, data analysis, machine learning, computer vision, and natural language processing.

Chunhua Shen (Member, IEEE) was with the computer vision program at NICTA (National ICT Australia), Canberra Research Laboratory for about six years. From 2012 to 2016, he held an Australian Research Council Future Fellowship. He is currently a Professor with the School of Computer Science, The University of Adelaide. He is also with Monash University, Australia. His research interests are in the intersection of computer vision and statistical machine learning.

Yanning Zhang (Senior Member, IEEE) is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. She is also the Organization Chair of the Ninth Asian Conference on Computer Vision (ACCV2009). Her research work focuses on signal and image processing, computer vision, and pattern recognition. She has published more than 200 papers in international journals, conferences, and Chinese key journals.