IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2019, accepted December 18, 2019, date of publication December 23, 2019, date of current version January 3, 2020.

*Digital Object Identifier 10.1109/ACCESS.2019.2961744*

# Robust License Plate Recognition With Shared Adversarial Training Network

**SHENG ZHANG[1], GUOZHI TANG[1], YULIANG LIU[1], AND HUIYUN MAO[2]**
[1]College of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China
[2]College of Computer Science, South China University of Technology, Guangzhou 510640, China

Corresponding author: Huiyun Mao (cshymao@scut.edu.cn)

This work was supported in part by NSFC under Grant 61936003, in part by the National Key Research and Development Program of China under Grant 2016YFB1001405, in part by Natural Science Foundation of Guangdong Province (GD-NSF) under Grant 2017A030312006, in part by the Guangdong Intellectual Property Office Project under Grant 2018-10-1, and in part by Guangzhou Science, Technology and Innovation Commission (GZSTP) under Grant 201704020134.

**ABSTRACT** Recently, deep learning has greatly promoted the performance of license plate recognition (LPR) by learning robust features from numerous labeled data. However, the large variation of wild license plates across complicated environments and perspectives is still a huge challenge to the robust LPR. To solve the problem, we propose an effective and efficient shared adversarial training network (SATN) in this paper, which can learn the environment-independent and perspective-free semantic features from wild license plates with the prior knowledge of standard stencil-rendered license plates, as standard stencil-rendered license plates are independent of complicated environments and various perspectives. Besides, to correct the features of heavily perspective distorted license plates perfectly, we further propose a novel dual attention transformation (DAT) module in the shared adversarial training network. Comprehensive experiments on AOLP-RP and CCPD benchmarks show that the proposed method outperforms state-of-the-art methods by a large margin on the LPR task.

**INDEX TERMS** Deep learning, license plate recognition (LPR), dual attention transformation (DAT), shared adversarial training network (SATN).

## I. INTRODUCTION

With the rapid development of intelligent transportation systems, license plate recognition (LPR) has attracted increasing research interests. It owns various potential applications, such as security and traffic control, vehicle re-identification [1], [2], and outdoor scene understanding [3], [4]. Much work has been done on the topic of LPR.

However, most of the existing methods require complex hardware to capture high-quality images, and others demand vehicles to pass a fixed access gate slowly or even at a full stop. It is still a challenging task to recognize license plates accurately in the wild due to the variations that suffer from appearance, blurring, noise, perspective, and illumination etc..

In the past few years, due to the powerful feature learning capabilities, convolutional neural networks (CNNs) have made significant progress in many computer vision tasks, such as object detection [5], semantic segmentation [2]. Thus, CNN-based LPR algorithms are also widely developed to solve the problem of recognizing license plates captured directly from the wild, for instance, transforming license plate recognition into a semantic segmentation task with the counting network to deal with appearance variations [6]. Although many LPR algorithms have been proposed [6], [7], they are still incapable of learning all variations in the wild. Consequently, their methods factually assume the input are high-quality images. Typically, the appearance of the license plates captured in wild scenes might suffer from the above degradations, causing deterioration in LPR performance. Hence, developing a robust LPR framework is highly indispensable, especially for wild scenes.

As is well-known, standard stencil-rendered license plates are free of distortions caused by environment and perspective, which present more semantic contents of license plates. While wild license plates contain various perspective distortion and complicated appearances, as shown in Fig. 1, which evidently throw a bad influence on the recognition of license plates. Extensive experiments have proven that the high-quality license plates can be more easily recognized

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

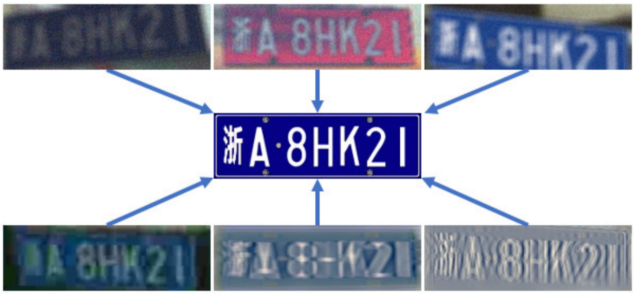

**FIGURE 1.** All different car license plates can be representable in standard stencil-rendered license plate which is free of numerous variations.

without the impact of complicated environment and various perspectives. In fact, standard stencil-rendered license plates are free of environment and perspectives, so that they can be exploited as prior information to guide model to recognize low-quality license plates as the pupils learn to read from a textbook. Consequently, it is possible to improve the performance of LPR by taking advantage of the semantic features which are environment-independent and perspective-free.

Besides, geometry correction for the heavy perspective distortions of wild license plates is indispensable. Although spatial transformer network (STN) [8] is well-known for correcting various affine and perspective transformations resulted from different perspectives, the simple localization network in STN always has limited receptive field. While the complicated localization network in STN is hard to be optimized because of the weak-supervision from the indirect losses of other subnetworks. Therefore, effective and efficient feature extraction module in STN is needed to improve the performance. Motivated by ECA-Net [9] and DAN [10], we introduce the channel attention and spatial attention modules in STN to form the dual attention transformation module (DAT), which could extract the wild license plate features efficiently and effectively to model the geometry transformations more robustly.

Meanwhile, apart from various perspective distortions, there also exist other distortions resulted from complicated environment, such as blurring, noise, and illumination. . ., which also have a bad influence on the LPR performance. To learn environment-independent and perspective-free semantic features of wild license plates simultaneously, we propose a shared adversarial training network (SATN) with the prior knowledge of standard stencil-rendered license plates, and embed the DAT module into SATN, which imposes a strong-supervision on the DAT module inversely. Actually, SATN is inspired by the generative adversarial networks (GANs) [11] which have demonstrated to be an extremely powerful tool for realistic image generation. GANs consist of a generator ($G$) and a discriminator ($D$). The discriminator can guide the generator to transfer a complicated data distribution to another specific distribution by adversarial training. When GAN is operated on a handwritten character set, such as MNIST dataset [12], it is exciting to observe that the generator could transfer noise vectors to realistic character images [13], [14]. Besides, [15] used GAN to generate images from street view house number (SVHN) dataset to the samples of MNIST dataset. Accordingly, as shown in Fig. 2(a), we argue that the generated license plates with wild license plates can be much similar to the stencil-rendered license plates by the GAN.

Specifically, SATN can make full use of the merits of discriminative model and generative model for LPR in Fig. 2(b). The DAT module is firstly exploited to correct various perspective distortions of the wild license plates. And then the feature encoder is used to encode features of wild and stencil-rendered license plates. Subsequently, the discriminator judges whether the encoded features come from wild or stencil-rendered license plates. With the prior knowledge generated by stencil-rendered license plates, it can instruct the feature encoder to distill environment-independent semantic features from wild license plates automatically. Finally, the encoded features are fed into the recognizer to recognize the wild license plates. The DAT module, feature encoder, discriminator, and the recognizer are alternately optimized by shared adversarial training. In the process of adversarial training, we use the feature activation loss to enhance the similarity of wild and stencil-rendered license plate features, which simultaneously provides a strong-supervision on the DAT module. Thus, we could get better performance of LPR.

Our main contributions are three-folds:

- we propose a novel dual attention transformation (DAT) module to correct the features of perspectively distorted wild license plates.
- To make the model learn environment-independent and perspective-free semantic features effectively and efficiently, we put forward a shared adversarial training network (SATN) with the prior knowledge of standard stencil-rendered license plates.
- Our proposed method outperforms previous state-of-the-art methods by a large margin on the AOLP-RP and CCPD benchmarks.

The rest of this paper is organized as follows: Section 2 reviews some related works. Section 3 formally introduces the proposed method detailedly. Section 4 presents the experimental results and analysis. Section 5 shows the conclusion and future work.

## II. RELATED WORKS

In the section, we will review some recovery methods for the degradation images and previous license plate recognition methods which are related to this work.

### A. SPATIAL TRANSFORMATION NETWORK

**STN** [8], namely spatial transformation network, is a solution to integrate learnable image warping into a neural network. A spatial transformer contains a subnetwork predicting a set of warp parameters followed by a *differentiable* warp function. STN has been proved effective in resolving geometric

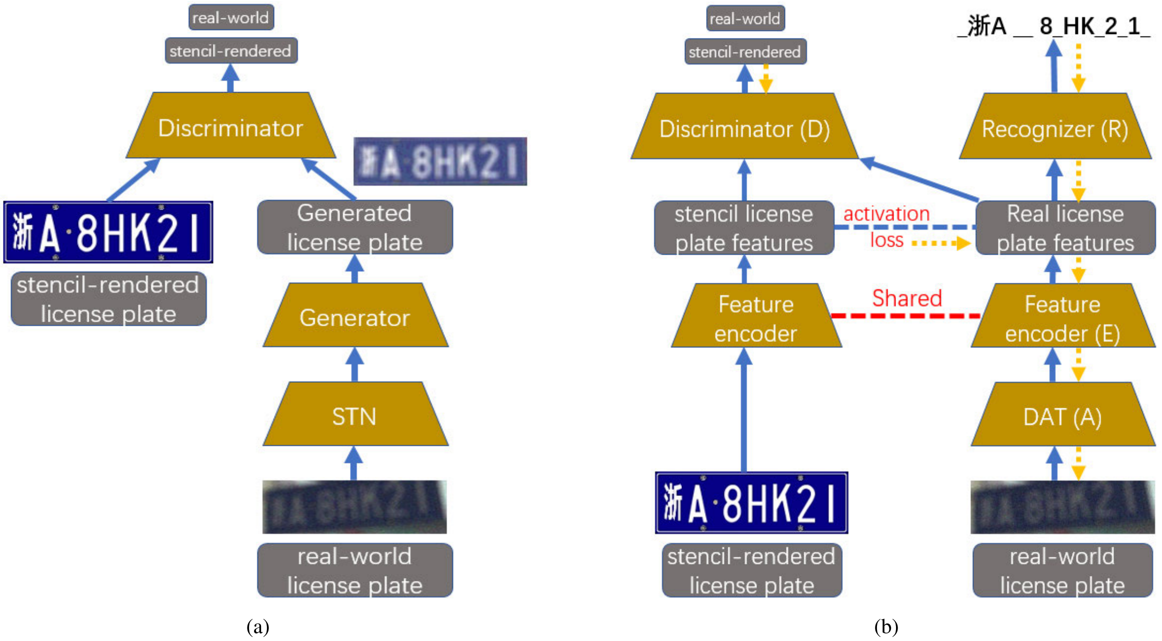

**FIGURE 2.** (a) Our usage of the GAN and STN in theory to generate high-quality license plates; (b) Our proposed SATN to recognize the wild license plates, solid blue arrows are for the forward-propagation, and dashed yellow arrows stand for the backward-propagation.

variations for classification tasks as well as other applications, such as image synthesis [16], scene text recognition [17]. In this paper, we improve the localization network in STN with channel and spatial dual attentions to correct the geometry variations of wild license plates.

### B. GENERATIVE ADVERSARIAL NETWORK

**GAN** is a class of generative models which are trained by optimizing a minimax objective between a generator ($G$) and a discriminator ($D$). Through the adversarial training, GAN has been testified to be able to learn a generative distribution that matches the expected distribution of a given data collection. One strength of GAN is that the loss function is essentially optimized by the discriminator network, which allows for training in cases where ground truth data with strong supervision is not available. In our work, we take advantage of the variant of GAN to guide the model to learn environment-independent and perspective-free semantic features from wild license plates simultaneously.

### C. LICENSE PLATE RECOGNITION

With the emerging of deep convolutional neural networks, numerous methods for LPR are proposed. Li and Shen [18] distill deep feature representations by using RNN to acquire sequential features of the license plate. Bulan *et al.* [19] evaluate domain shifts between target and several source domains for choosing a domain which outputs the best recognition performance based on fully convolutional network [2]. However, these methods only consider high-quality license plates except for low-quality ones, which may decrease the performance heavily in complex wild scenes. Besides, these methods make little or no effort to learn environment-independent and perspective-free semantic features of wild license plates, while possessing a high computational complexity. In this work, unlike existing methods, we adopt shared adversarial training to learn environment-independent and perspective-free semantic features of wild license plates in order to obtain high LPR performance. To the best of our knowledge, this is the first time that GAN is applied to LPR of complicated environment and heavy perspective distortion. Meanwhile, our method is not only effective but also efficient for real-time application of LPR.

## III. SHARED ADVERSARIAL TRAINING NETWORK

The architecture of the proposed model SATN is illustrated in Fig. 2(b), the performance of LPR is improved by learning perspective-free and environment-independent semantic features of wild license plates, where we provide standard stencil-rendered license plates as prior knowledge with synthetic method [20]. SATN consists of four neural network components: 1) Dual attention transformation module $A$ for perspective correction; 2) Feature encoder $E$ for extracting environment-independent semantic features; 3) Discriminator $D$ to classify the encoded features of wild and stencil-rendered license plates; 4) Recognizer $R$ to predict the license plate label. $A$, $E$, $D$, and $R$ are alternately optimized by the shared adversarial training algorithm.

### A. DUAL ATTENTION TRANSFORMATION

Geometry correction with STN is critical for the heavy perspective distortions of wild license plates. However, the

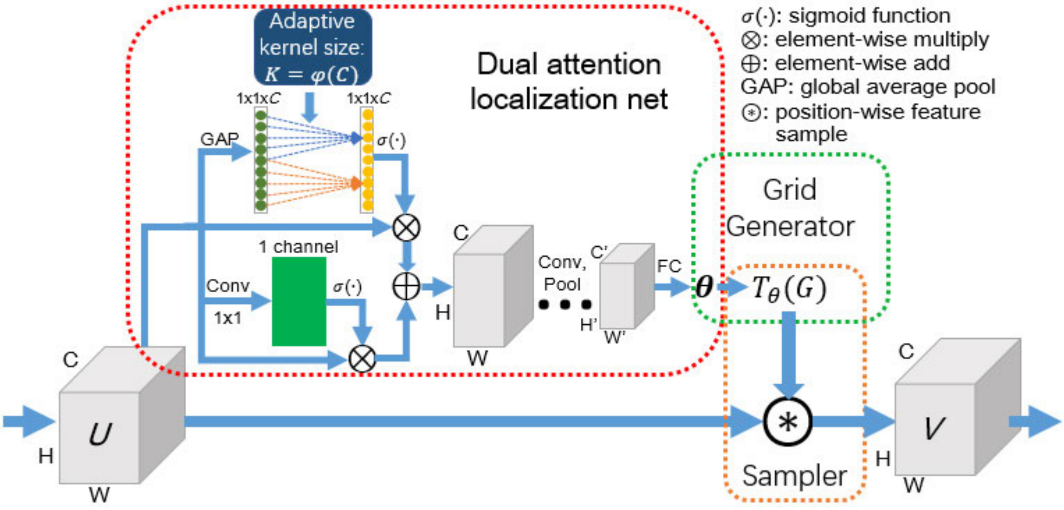

**FIGURE 3. Dual attention transformation.**

simple localization network in STN always makes it have limited receptive field, and the complicated localization network in STN is hard to be optimized because of the weak-supervision from the indirect losses of other subnetworks. Therefore, effective and efficient feature extraction module in STN will be indispensable to improve the performance. Recently, attention mechanism has been proven to offer great potential in improving the performance of feature extraction for CNN. However, most existing methods are dedicated to leveraging more sophisticated attention modules to achieve better performance, inevitably increasing the computational burden. In fact, LPR task demands both high performance and efficiency. Thus, we introduce two extremely light-weight attention modules in STN for boosting the performance of the localization network, namely, efficient channel attention (ECA) [9] and spatial attention (SA) [21], which forms the novel dual attention transformation module (DAT) in SATN. DAT can extract the broad context features efficiently and effectively to model the geometry transformations and correct the heavy perspective distortions of wild license plates. As displayed in Fig. 3, it consists of three components: a dual attention localization network, a grid generator, and a sampler. We mainly introduce the dual attention localization network, as the last two components are roughly the same as those in the work [8] except that grid height ($h = 4$) and width ($w = 16$) are specific for license plates.

### 1) DUAL ATTENTION LOCALIZATION NETWORK

#### *a: ECA*

By revisiting the channel attention module in SENet [22], ECA shows that avoiding dimensionality reduction and appropriate local cross-channel interaction are important for learning effective channel attention. As shown in Fig. 3, assuming the output of one convolution block is $U \in \mathbb{R}^{W\times H\times C}$, it firstly conducts the global average pooling (GAP) for $U$ by Equation 1.

$$y = GAP(U) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H} U_{ij} \tag{1}$$

Then, the weight of $y_i \in y$ can be calculated as Equation 2 by only considering the local interaction between each channel and its $k$ neighbors.

$$\omega_i = \sigma\left(\sum_{j=1}^{k}\alpha^j y_i^j\right), \quad y_i^j \in \Omega_i^k \tag{2}$$

where $\Omega_i^k$ indicates the set of $k$ adjacent channels of $y_i$, $\sigma$ is a sigmoid function, $\alpha^j$ is one parameter of a shared 1D convolution kernel of size $k$. Since it is expected that larger size of channels favors long-range interaction while smaller size of channels prefers short-term interaction, $k$ should be adaptively determined by channels $C$ as in Equation 3.

$$k = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{3}$$

where $|t|_{odd}$ indicates the nearest odd number of $t$. In this paper, we set $\gamma$ and $b$ to 2 and 1, respectively. Obviously, ECA can bring significant performance gain with only $k$ ($k \le 9$) parameters.

#### *b: SPATIAL ATTENTION*

Spatial attention is achieved through a convolution $q = W_{sa} \star U$ with weight $W_{sa} \in R^{1\times C\times 1\times 1}$, generating a projection tensor $q \in R^{H\times W}$. Each $q_{ij}$ of the projection represents the linearly combined representation for all channels $C$ for a spatial location $(i, j)$. This projection is passed through a sigmoid layer $\sigma(\cdot)$ to obtain the spatial attention weights. Theoretically, spatial attention assigns more importance to relevant spatial location features and ignores irrelevant ones, which can also improve the performance effectively.

Obviously, the proposed dual attention transformation module is extremely lightweight to correct the perspective distortions of wild license plates, and its effectiveness is noticeable in Table 6.

### B. FEATURE ENCODING, DISCRIMINATING, RECOGNITION

Although the aforementioned novel DAT module can solve the heavy perspective distortions of wild license plates, it needs the strong-supervision information to exert its potential. Meanwhile, the license plates captured from the wild scenes are also degraded by the sophisticated environment factors including high exposure, blurring, occlusion, unbalanced illumination, and inherent plate damage, . . . To extract the environment-independent and perspective-free semantic features simultaneously, we embed the DAT module and recognition branch into the well-known GAN [11] and make them share some backbone layers, which forms the shared adversarial training network (SATN). As we know, GAN has been demonstrated to be an extremely powerful tool for realistic image generation. It consists of a generator ($G$) and a discriminator ($D$). The discriminator can guide the generator to transfer a complicated data distribution to another specific distribution by adversarial training. Thus, we expect that the generated semantic features of wild license plates can be much similar to those of the stencil-rendered license plates by the SATN, since stencil-rendered license plates are free of the impact of the complicated environment and heavy perspective distortions.

Specifically, assuming the outputs of the DAT module for wild and stencil-rendered license plates are $V_r$ and $V_s$, respectively. The corresponding feature maps are denoted as $E(V_r, \theta_e)$ and $E(V_s, \theta_e)$ which are encoded by the feature encoder $E$ in SATN (Note: $\theta_e$ contains the learnable parameters of DAT module hereafter). We calculate the following smooth L1 loss between the features $E(V_r, \theta_e)$ and $E(V_s, \theta_e)$ like the generative model objective in [23],

$$SL_1(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geq 1 \end{cases} \tag{4}$$

which forces the learned feature $E(V_r, \theta_e)$ to be similar to the feature $E(V_s, \theta_e)$ as much as possible.

The discriminator $D$ is used to judge whether the encoded features by $E$ come from wild or stencil-rendered license plates. We denote the feature label as $y_d$. The features $E(V_r, \theta_e)$ are assigned a label of $y_d = 1$, while the features $E(V_s, \theta_e)$ correspond to $y_d = 0$. The probability of $y_d = 1$ and $y_d = 0$ are expressed as Equations 5 and 6, respectively:

$$P(y_d = 1|E(V_r, \theta_e); \theta_d) = \frac{1}{1 + e^{-D[E(V_r,\theta_e),\theta_d]}} \tag{5}$$

$$P(y_d = 0|E(V_s, \theta_e); \theta_d) = 1 - P(y_d = 1|E(V_r, \theta_e); \theta_d) \tag{6}$$

where $\theta_d$ denotes the parameters of $D$. $D$ is optimized by minimizing the discriminator loss $L_d(E(V_s,\theta_e), E(V_r,\theta_e); \theta_d)$ as:

$$\begin{aligned} &L_d(E(V_s, \theta_e), E(V_r, \theta_e); \theta_d) \\ &\quad = \begin{cases} -log(P(y_d = 0|E(V_s, \theta_e); \theta_d)) \\ -log(P(y_d = 1|E(V_r, \theta_e); \theta_d)) \end{cases} \end{aligned} \tag{7}$$

Furthermore, only features $E(V_r, \theta_e)$ of wild license plates are sent into the recognizer $R$ for sequence recognition. Specifically, a series of convolution and pool operations are applied on $E(V_r, \theta_e)$ to obtain a unit height feature maps $F_r$; Then, we leverage two bidirectional LSTM layers [24] on $F_r$ to acquire the feature maps $F_p$, which encodes the context sequence information and predicts the license plate labels. Finally, a CTC loss [25] $L_r$ for the alignment of the features $F_p$ and the target license plate labels is evaluated. Noticeably, we do not use the attention decoder in recognizer $R$ because CTC is more efficient and enough for the LPR task.

In a word, the proposed SATN is learnt to discover environment-independent and perspective-free semantic features that can fool $D$. The encoded semantic features can be viewed as underlying features which stencil-rendered and wild license plates share in common, $D$ can not distinguish them. Namely, the more environment-independent and perspective-free features SATN extracts, the larger discriminator loss $L_d(E(V_s, \theta_e), E(V_r, \theta_e); \theta_d)$ is.

Therefore, the proposed SATN model can be optimized by minimizing the adversarial train loss $L(I_s, I_r; \theta_d, \theta_e, \theta_r)$ expressed as:

$$L = \lambda SL_1 - \alpha L_d + \beta L_r \tag{8}$$

where $\lambda, \alpha$, and $\beta$ are the hyper-parameters to control the trade-off between the smooth L1 loss, discriminator loss, and recognition loss, and $\lambda = 10, \alpha = \beta = 1$ empirically. $I_s, I_r$ stand for the stencil-rendered and wild license plates, respectively.

### C. NETWORKS DESIGN

In this subsection, we introduce all architectures of the components DAT, $E$, $D$ and $R$ in the proposed SATN model detailedly.

The architecture of the dual attention transformation (DAT) module is elaborated in Table 1. Besides, it is in front of the encoder $E$ and recognizer $R$, and shared by them.

The architectures of the feature encoder $E$, recognizer $R$ and discriminator $D$ are elaborated in Table 2. Each convolutional layer is followed by a batch normalization layer and a ReLU layer.

### D. OPTIMIZATION

SATN model is optimized alternately for robust wild license plate recognition. *DAT* is trained to correct the perspective distortions of wild license plates by the strong supervision of the smooth L1 loss, $D$ is learnt to distinguish the features of standard stencil-rendered license plate $I_s$ from the features of wild license plate $I_r$, $E$ is optimized to encode environment-independent discriminative features of $I_r$ that

**TABLE 1. The architecture of the dual attention transformation (DAT) module. "Owned" means shared by *E* and *R*, k, s, p are kernel, stride and padding sizes, respectively. For example, *s*2 represents a 2 stride size. "$*$" stands for module with dropout layer (drop ratio = 0.3).**

| Dual attention transformation (DAT) | | |
|---|---|---|
| Type | Configurations | Size |
| Input | - | 3x32x128 |
| Convolution | maps:10, k3, s1, p1 | 10x32x128 |
| MaxPooling | k2, s2 | 10x16x64 |
| Convolution$*$ | maps:64, k3, s1, p1 | 64x16x64 |
| MaxPooling | k2, s2 | 64x8x32 |
| Convolution$*$ | maps:20, k3, s1, p1 | 20x8x32 |
| MaxPooling | k2, s2 | 20x4x16 |
| Resize | - | 1280 |
| FC | maps:64 | 128 |
| Tanh | - | 128 |

**TABLE 2. Architectures of the feature encoder *E* and recognizer *R*. "Owned" means shared by *E* and *R*, k, s, p are kernel, stride and padding sizes, respectively. For example, *s*2 × 1 represents a 2 × 1 stride size. "$*$" stands for module with dropout layer (*drop_ratio* = 0.3).**

| Encoder $E$ and Recognizer $R$ | | | |
|---|---|---|---|
| Type | Configurations | Size | Owned |
| Input | - | 3x32x128 | $E$,$R$ |
| Convolution | maps:64, k3, s1, p1 | 64x32x128 | $E$,$R$ |
| MaxPooling | k2, s2 | 64x16x64 | $E$,$R$ |
| Convolution | maps:128, k3, s1, p1 | 128x16x64 | $E$,$R$ |
| MaxPooling | k2, s2 | 128x8x32 | $E$,$R$ |
| Convolution | maps:256, k3, s1, p1 | 256x8x32 | $E$,$R$ |
| Convolution | maps:256, k3, s1, p1 | 256x8x32 | $E$,$R$ |
| MaxPooling | k2, s2 | 256x4x16 | $R$ |
| Convolution | maps:512, k3, s1, p1 | 512x4x16 | $R$ |
| Convolution | maps:512, k3, s1, p1 | 512x4x16 | $R$ |
| MaxPooling | k2, s2x1 | 512x2x16 | $R$ |
| Convolution | maps:512, k2, s1 | 512x1x15 | $R$ |
| BLSTM$*$ | hidden unit:256 | 256x1x15 | $R$ |
| BLSTM$*$ | hidden unit:256 | 256x1x15 | $R$ |

| Discriminator $D$ | | |
|---|---|---|
| Type | Configurations | Size |
| Input | $E(V_s, \theta_e)$, $E(V_r, \theta_e)$ | 256x4x16 |
| Convolution$*$ | maps:256, k3, s1, p1 | 256x4x16 |
| MaxPooling | k2, s2 | 256x2x8 |
| Convolution$*$ | maps:64, k3, s1, p1 | 64x2x8 |
| Convolution$*$ | maps:64, k3, s1, p1 | 64x2x8 |
| UpSample | k2, s2 | 64x4x16 |
| Convolution | maps:1, k3, s1, p1 | 1x4x16 |

can fool $D$, and $R$ is optimized to predict the wild license plate labels from the encoded features. With $D$ being more powerful in distinguishing whether features are environment-independent or not and $R$ being more precise in recognizing plate labels, *DAT* and $E$ strive to learn the perspective-free and environment-independent features to compete with D. Finally, DAT, $E$, $D$ and $R$ improve each other in the progress of shared adversarial training. In other words, $D$ is optimized by minimizing the discriminator loss $L_d$, while the perspective-free and environment-independent features encoded by *DAT* and $E$ respectively, which are similar to the features encoded from stencil-rendered license plates, will result in maximizing the discriminator loss $L_d$. I.e., *DAT*, $E$,

**Algorithm 1** Shared Adversarial Training Algorithm

**Input**: Wild and standard license plate ($I_r$, $I_s$), learn rate $\mu$.
**Parameters**: Randomly initialized $\theta_e$, $\theta_d$, $\theta_r$.
**Output**: The optimal parameters $\hat{\theta}_e$, $\hat{\theta}_d$, $\hat{\theta}_r$.
**Start training**:

```
// Pretrain a whole license plate recognizer
For number of pretraining epochs do
    For number of Mini-batches do
        $(\theta_e,\theta_r)\leftarrow(\theta_e,\theta_r)-\mu\left(\frac{\partial SL_1(I_s,I_r;\theta_e)+L_r(I_r;\theta_e,\theta_r)}{\partial\theta_e},\frac{\partial L_r(I_r;\theta_e,\theta_r)}{\partial\theta_r}\right)$
    End for
End for
// the adversarial training
Repeat
    For number of training epochs do
        For number of Mini-batches do
            // for discriminator
            $\theta_d \leftarrow \theta_d - \mu\alpha\frac{\partial L_d(E(V_s,\theta_e),E(V_r,\theta_e))}{\partial\theta_d}$
            // for feature encoder
            $\theta_e \leftarrow \theta_e - \mu\lambda\left(\frac{\partial SL_1(I_s,I_r;\theta_e)}{\partial\theta_e}\right)$
            // for recognizer
            $(\theta_e,\theta_r)\leftarrow(\theta_e,\theta_r)-\mu\beta\left(\frac{\partial L_r(I_r;\theta_e,\theta_r)}{\partial\theta_e},\frac{\partial L_r(I_r;\theta_e,\theta_r)}{\partial\theta_r}\right)$
        End for
    End for
Until convergence
$\hat{\theta}_e = \theta_e, \hat{\theta}_r = \theta_r, \hat{\theta}_d = \theta_d$
return $\hat{\theta}_e, \hat{\theta}_r, \hat{\theta}_d$
```

$D$ and $R$ play the *minimax* game with loss function:

$$\begin{aligned}\min_{\theta_e,\theta_r}\max_{\theta_d}\ & L\,(\theta_e, \theta_d, \theta_r)\\ &= \lambda SL_1\,(I_r, I_s; \theta_e)\\ &\quad - \alpha L_d(E(V_s, \theta_e), E(V_r, \theta_e); \theta_d) + \beta L_r\,(I_r; \theta_e, \theta_r) \qquad (9)\end{aligned}$$

Suppose the optimal parameters are $\hat{\theta}_e$,$\hat{\theta}_d$,$\hat{\theta}_r$, then we have

$$\hat{\theta}_e, \hat{\theta}_r = \text{argmin}\{SL_1(\theta_e) + L_r(\theta_e, \theta_r, \hat{\theta}_d)\} \qquad (10)$$

$$\hat{\theta}_d = \text{argmax}\{L_d(E(V_s, \hat{\theta}_e), E(V_r, \hat{\theta}_e); \theta_d)\} \qquad (11)$$

The contradictory optimization targets for the parameters $\theta_e$ in $E$ and $\theta_d$ in $D$ make it hard for learning $E$, $D$ and $R$ in one updating step. Thus, we constrain the update to the respectively specific component in SATN model. DAT, $E$, $D$ and $R$ are alternatively optimized following the adversarial learning framework. As shown in Algorithm 1, we link DAT, $E$ and $R$ to pretrain a primary license plate recognizer. Afterwards, we alternatively optimize the parameters of DAT, $E$, $D$ and $R$ to fine-tune the model.

## IV. EXPERIMENT

In the section, we introduce a series of benchmarks and experimental details for the proposed method SATN.

### A. BENCHMARKS

**CCPD** [20] is the largest publicly available diverse license plate (LP) benchmark to date with over 250k unique car

**TABLE 3.** Illustrations of different sub-sets in CCPD.

| Sub-sets | Illustration | train/test |
|---|---|---|
| Base | The only common feature of these is the inclusion of a license plate. | 100k/100k |
| DB | Illuminations on the LP area are dark, uneven or extremely bright. | 10k/10k |
| FN | The distance from the LP to the shooting location is relatively far or near. | 10k/10k |
| Rotate | Great horizontal tilt degree $(20^\circ, 50^\circ)$ and the vertical tilt degree $(-10^\circ, 10^\circ)$ | 5k/5k |
| Tilt | Great horizontal tilt degree $(15^\circ, 45^\circ)$ and the vertical tilt degree $(15^\circ, 45^\circ)$ | 5k/5k |
| Blur | Blurry largely due to hand jitter while taking pictures. | 2.5k/2.5k |
| Weather | Images taken on a rainy day, snow day or fog day. | 5k/5k |
| Challenge | The most challenging images for LPDR to date. | 5k/5k |

**TABLE 4.** LPR precision (percentage) on each test subset of CCPD. HC denotes Holistic-CNN recognition [30]. AP: average precision.

| Methods | FPS | AP | Base | DB | FN | Rotate | Tilt | Weather | Challenge |
|---|---|---|---|---|---|---|---|---|---|
| Cascade classifier+HC [30] | 29 | 58.9 | 69.7 | 67.2 | 69.7 | 0.1 | 3.1 | 52.3 | 30.9 |
| SSD300+HC [31] | 35 | 95.2 | 98.3 | 96.6 | 95.9 | 88.4 | 91.5 | 87.3 | 83.8 |
| YOLO9000+HC [32] | 36 | 93.7 | 98.1 | 96.0 | 88.2 | 84.5 | 88.5 | 87.0 | 80.5 |
| Faster-RCNN+HC [5] | 13 | 92.8 | 97.2 | 94.4 | 90.9 | 82.9 | 87.3 | 85.5 | 76.3 |
| TE2E [33] | 3 | 94.4 | 97.8 | 94.8 | 94.5 | 87.9 | 92.1 | 86.8 | 81.2 |
| RPnet [20] | 61 | 95.5 | 98.5 | 96.9 | 94.3 | 90.8 | 92.5 | 87.9 | 85.1 |
| Our SATN | **113** | **98.9** | **99** | **99.4** | **99.3** | **99.4** | **99.4** | **99.3** | **96.5** |

images, and it is divided into two equal parts of training set and testing set. Each LP number is comprised of a Chinese character, an English letter, and five English letters or numbers, which results in the 69 classes of all license plates (Note: including one "blank" class and one "unknown" class). Besides, it consists of 8 sub-sets, which are captured from different complicated environments and various perspectives. The distribution and illustrations of all sub-sets in CCPD are shown in Table 3.

**AOLP-RP** [26] includes 611 license plate images gathered in Taiwan, each image contains ten numbers and 25 letters (except character "O"). The dataset is very challenging because the angles of some license plates contain heavy perspective distortions. What is more, in terms of resolution, all captured images are relatively simple because they own high-resolution rather than other datasets.

### B. IMPLEMENTATION DETAILS

All the reported experimental results are built on the Pytorch framework, and our method has done on one NVIDIA TITAN X GPU and one Intel Core i7-6700K CPU. In all the experiments, we use a gradient clipping trick and the Adadelta optimizer [27] with momentum value 0.9. The proposed network is trained in roughly 200 epoches with a batch size of 32. The weights in discriminator $D$ and DAT are initialized from a zero-mean Gaussian distribution with a standard deviation of 0.02, and the remaining weights in SATN are initialized by the method in [28]. The initial learning rate is 0.1 for the DAT module, encoder $E$ and recognizer $R$, and $10^{-3}$ for the discriminator $D$.

### C. COMPARISON

In this subsection, we compare our method with previous state-of-the-art methods of LPR. As seen in Table 4, obviously, our method outperforms the prior best method by a large margin on the average precision (AP) metric. More importantly, on the subsets of "`DB`, `FN`, `Rotate`, `Tilt`, `Weather`, `Challenge`" of CCPD benchmark except "`Base`" subset, the performance of our method is significantly better than existing methods, which strongly certificates that our proposed SATN model can learn the perspective-free and environment-independent features of wild license plates. Additionally, in Table 5, we also evaluate our method on another benchmark AOLP-RP which suffers from the perspective distortion heavily, and it still performs best.

**TABLE 5.** LPR precision (percentage) on AOLP-RP.

| Methods | accuracy (%) |
|---|---|
| [26] | 85.76 |
| [33] | 88.38 |
| [7] | 98.36 |
| [6] | 99.02 |
| [34] | 98.85 |
| Our SATN | **99.67** |

**TABLE 6.** The effectiveness of each component in our SATN model on the CCPD benchmark. "SAT" in the above table stands for shared adversarial training, "*R*" is baseline recognition, "√" means including certain component.

| Methods | DAT | SAT | AP | Base | DB | FN | Rotate | Tilt | Weather | Challenge |
|---|---|---|---|---|---|---|---|---|---|---|
| CRNN | | | 90.5 | 92.4 | 92.8 | 93.1 | 87.9 | 88.4 | 91.3 | 87.6 |
| CRNN+DAT | √ | | 93.1 | 94.8 | 95 | 96.2 | 93.9 | 94.3 | 93.5 | 92.8 |
| CRNN+SAT | | √ | 96.4 | 97.9 | 98.4 | 98.7 | 94.2 | 95.1 | 97.3 | 93.2 |
| SATN | √ | √ | **98.9** | **99** | **99.4** | **99.3** | **99.4** | **99.4** | **99.3** | **96.5** |

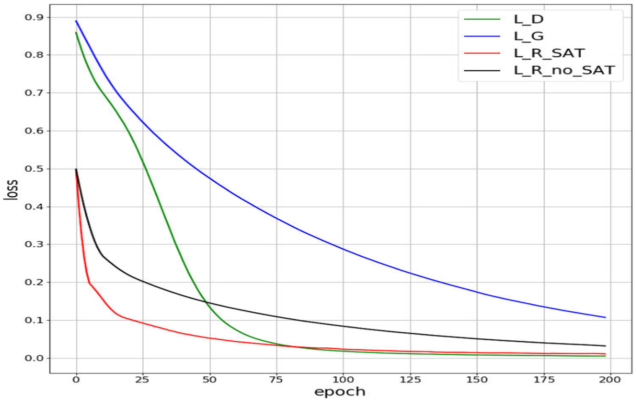

**FIGURE 4.** The visualization of the losses of Shared Adversarial Training. "L_R, L_D, L_G" stand for the losses of recognition, discrimination and encoding feature, respectively. The values of "L_G" are scaled by a factor of 0.1 to draw the figure better.

To make the conclusion much more convincing, some ablation experiments are conducted in Table 6 and Figures. 4, 5, 6. In Table 6, the baseline is CRNN [29]. With our DAT module in "CRNN" model, the promotions of performance on "`Rotate`, `Tilt`" subsets are evidently far larger than those of the remaining subsets of CCPD, which proves that our DAT module can correct the perspective distortions effectively. Meanwhile, the transformed results in Fig. 6 also show that

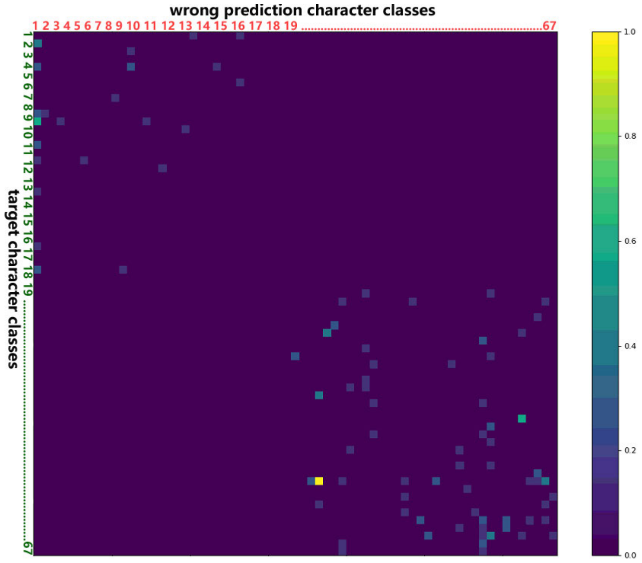

**FIGURE 5.** The confusion matrix of the target and wrong prediction character classes. The vertical axis are target classes, and the horizontal axis are the wrong prediction character classes.

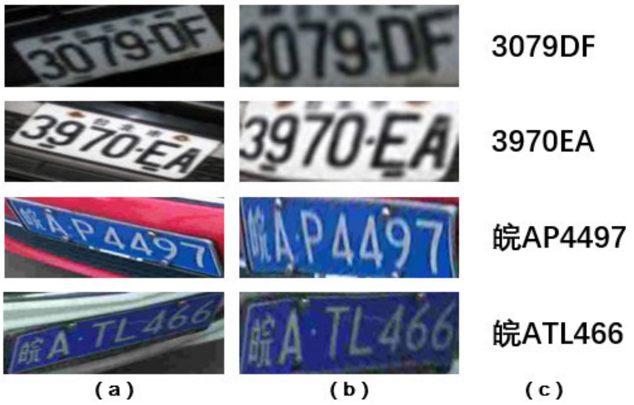

**FIGURE 6.** The results of the proposed DAT module for the perspective distortions of wild license plates. (a) Input license plates; (b) Transformed results by DAT; (c) Recognition results.

the proposed DAT module is vital for the recognition of perspective distortion license plates. Further, via using SAT in "CRNN" model, the environment-independent features can be learned effectively to promote the performance substantially. While we visualize the training process in Fig. 4, drastically, the red line with SAT not only converges much faster than the black line without SAT but also exhibits lower training losses, which shows the powerful ability of our SAT.

In Fig. 5, we analyze the erroneous recognition results via a confusion matrix of the target and wrong prediction character classes, and the color intensity in the right-side bar describes the error level of the target character to be predicted as another character. Through Fig. 5, we observe that: (1) Class imbalance will result in the recognition error in some way, for instance, some abbreviations of other provincial capitals are wrongly predicted as the first class due to a lot of training samples of the first class; (2) Appearance similarity is another reason, for example, the highest error rate is target character class "D" which is erroneously recognized as class "0" because the font type used on the license plates makes them appear similar; (3) Most of the wrong prediction classes concentrate on the English characters and numbers which are easily affected by the complicated environment because of less strokes.

## V. CONCLUSION

In this paper, we propose an effective and efficient shared adversarial training network (SATN), which can learn the environment-independent and perspective-free features from wild license plates with the prior knowledge of standard stencil-rendered license plates, as standard stencil-rendered license plates are independent of complicated environment and various perspectives. Besides, to correct the features of perspectively distorted license plates perfectly, we further propose a novel dual attention transformation (DAT) module in the shared adversarial training network. Comprehensive experiments on AOLP-RP and CCPD benchmarks show that the proposed method outperforms state-of-the-art methods by a large margin.

## ACKNOWLEDGMENT

The authors would like to thank Prof. L. Jin for his helpful discussion and advice for this research project.

## REFERENCES

[1] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2167–2175.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[3] S. Cherng, C. Y. Fang, C. P. Chen, and S. W. Chen, "Critical motion detection of nearby moving vehicles in a vision-based driver-assistance system," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 70–82, Mar. 2009.

[4] S. Noh and M. Jeon, "A new framework for background subtraction using multiple cues," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 493–506.

[5] 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.

[6] J. Zhuang, S. Hou, Z. Wang, and Z.-J. Zha, "Towards human-level license plate recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 306–321.

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

[8] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[9] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," 2019, *arXiv:1910.03151*. [Online]. Available: https://arxiv.org/abs/1910.03151

[10] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[12] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," [Online]. Available: http://yann.lecun.com/exdb/mnist

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[15] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, *arXiv:1611.02200*. [Online]. Available: https://arxiv.org/abs/1611.02200

[16] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky, "Deepwarp: Photorealistic image resynthesis for gaze manipulation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 311–326.

[17] F. Zhan and S. Lu, "ESIR: End-to-end scene text recognition via iterative image rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2059–2068.

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

[19] 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.

[20] 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)*, 2018, pp. 255–271.

[21] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 421–429.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[23] 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.*, Oct. 2017, pp. 2223–2232.

[24] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. (IJCNN) Neural Comput., New Challenges Perspect. New Millennium*, vol. 3, Jul. 2000, pp. 189–194.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[26] 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.

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

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[29] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[30] J. Špaňhel, J. Sochor, R. Juránek, A. Herout, L. Maršík, and P. Zemčík, "Holistic recognition of low quality license plates by CNN using track annotated data," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug./Sep. 2017, pp. 1–6.

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

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

[33] 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. 2018.

[34] Y. Lee, J. Lee, H. Ahn, and M. Jeon, "SNIDER: Single noisy image denoising and rectification for improving license plate recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2019, pp. 1–10.

**SHENG ZHANG** received the B.S. degree from Hohai University, Nanjing, China, in 2014. He is currently pursuing the Ph.D. degree with the DLVC Laboratory, South China University of Technology. His major research interests include object detection, text detection and recognition, deep learning, and image processing algorithms.

**GUOZHI TANG** received the bachelor's degree from Yunnan University, Yunnan, China, in 2019. He is currently pursuing the master's degree with the College of Electronic and Information Engineering, South China University of Technology, China. His major research interests include object detection, text detection, and information extraction.

**YULIANG LIU** received the B.S. degree from the South China University of Technology, in 2016, where he is currently pursuing the Ph.D. degree. His research interests include deep learning and object detection and text recognition.

**HUIYUN MAO** received the Ph.D. degree from the South China University of Technology (SCUT), Guangzhou, China, in 2011. She is currently a Lecturer with the College of Computer Science, South China University of Technology. She has authored over 20 scientific articles and patents. Her research interests include image processing, machine learning, and intelligent systems.

• • •