HomoNet: Unified License Plate Detection and Recognition in Complex Scenes

Yuxin Yang[1], Wei Xi[1], Chenkai Zhu[2], and Yihan Zhao[1(✉)]

[1] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
yangdx6@gmail.com, weixi.cs@gmail.com, 18292047025@163.com
[2] SenseTime Group Ltd., Shenzhen, China
zhuchenkai@sensetime.com

Abstract. Although there are many commercial systems for license plate detection and recognition (LPDR), existing approaches based on object detection and Optical Character Recognition (OCR) are difficult to achieve good performance in both efficiency and accuracy in complex scenes (e.g., varying viewpoint, light, weather condition, etc). To tackle this problem, this work proposed a unified end-to-end trainable fast perspective LPDR network named HomoNet for simultaneous detection and recognition of twisted license plates. Specifically, we state the homography pooling (HomoPooling) operation based on perspective transformation to rectify tilted license plates. License plate detection was replaced with keypoints location to obtain richer information and improve the speed and accuracy. Experiments show that our network outperforms the state-of-the-art methods on public datasets, such as 95.58%@22.5 ms on RP and 97.5%@19 ms on CCPD.

Keywords: License plate · Keypoints location · HomoPooling

1 Introduction

With the development of deep learning, license plate detection and recognition (LPDR) has been widely applied in intelligent transportation and surveillance systems. In these scenarios, the high accuracy and fast speed must be met at the same time. Besides, cars run at high speeds in different directions in complex situations, causing the license plate (LP) may not face the camera, making it to be twisted, rotated, perspective and distorted, and posing a challenge to LPDR. Many researches have worked to improve performance in recent years.

Most systems [9,11,15,25,26] divided the LPDR into two independent parts: LP detection and LP recognition. These methods have three limitations: (1) The

This work was supported by National Key R&D Program of China 2018AAA0100500, NSFC Grant No. 61772413, 61802299, and 61672424.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2021
Published by Springer Nature Switzerland AG 2021. All Rights Reserved
H. Gao et al. (Eds.): CollaborateCom 2020, LNICST 350, pp. 268–282, 2021.
https://doi.org/10.1007/978-3-030-67540-0_16

detection result is a bounding box that can only use two points to locate the LP position. In complex scenes with different views, the boundary of LP in video surveillance is generally not a rectangle, which leads to misalignment between the actual LP region and the extracted region by a bounding box, and results in some important information missing. (2) The neural network is difficult to recognize tilted characters. The bounding box only can crop the region of LP ignoring its perspective change because of the direction to the camera. Thus methods lacking correction can not efficiently solve the problem of recognizing twisted LP in complex scenes. (3) There is a correlation between the detection and recognition which was ignore in these articles. The two subtasks can not train end-to-end jointly and spend more time on the inference process. Thus, Li et al. [15] proposed a unified deep neural network TE2E, as shown in Fig. 1a. The lack of correction, however, leads to poor performance on rotating and tilt license plates. Another problem is the detection module is anchor-based which limits the whole net speed.

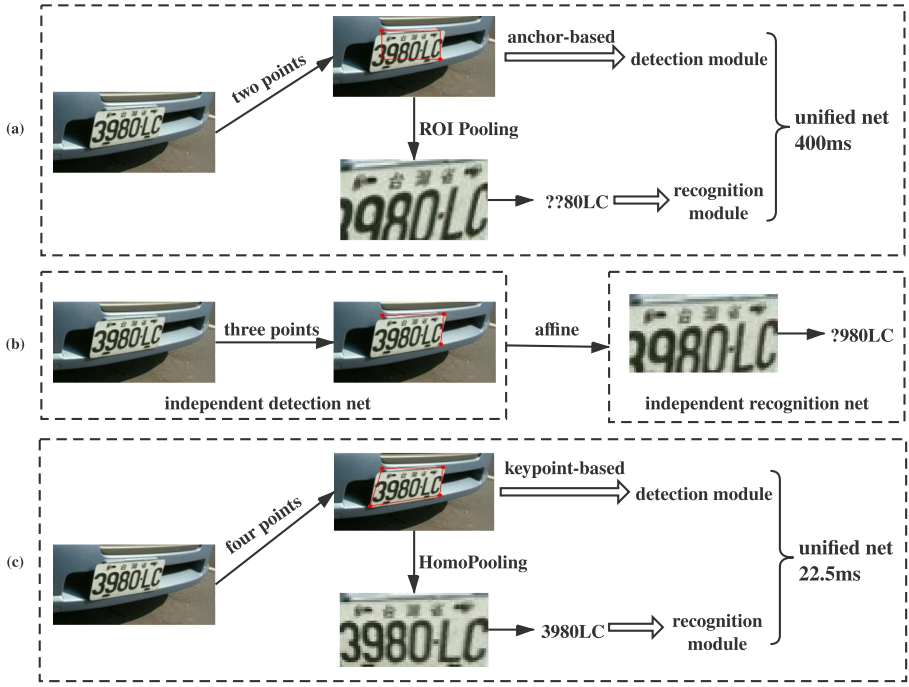

Fig. 1. Comparison between there methods: (a) TE2E, (b) warp-net, (c) HomoPooling.

Inspired from text spotting, affine transformation was introduced to solve the problem of LP distortion [10,15,18]. Affine transformation is equivalent to positioning three coordinates in the original picture, determining the transformation matrix, and then rectifying the license plate. This method can only improve the

recognition problem of perspective character to some extent, but cannot fundamentally solve it. For example, as shown in Fig. 1b, character 9 is recognized benefiting from the transformation but character 3 is still difficult to identify due to the detection error of the upper left corner. What's more, the affine transformation is not differentiable in warp-net [26] leading to the issue that we have stated above.

In this paper, based on perspective transformation, a unified network with HomoPooling operation was proposed (see Fig. 1c). Even if there is an error points, four points can still extract all valid information and correct the twisted LP to improve the accuracy of the recognition. Moreover, HomoNet puts the detection, rectification and recognition into one network to ensure end-to-end training, improving the efficiency and accuracy.

The method of LP detection mainly learned from the object detection task. Anchor-based methods need the network to predict massive bounding boxes and non-maximum-suppression (NMS) operations and this method takes up computing resources and slows down the detection time. At the same time, it is hard to get high location accuracy. Different from the most object detection tasks that one image has multiple targets, a car only has one license plate. Thus, to improve the accuracy and speed of detection, we directly predict the heat map to get the key points of the LP, aspired by human skeleton key point detection. Particularly, experiments have been done to prove it. Regressing the four offsets of each pixel relative to the ground truth as in FOTS [18], the root mean square error in Road Patrol (RP) dataset was 13.6, while the proposed method was 3.4. As shown in Fig. 1a and Fig. 1c, benefiting from the keypoint-based method, the whole network speed improved from 400 ms to 22.5 ms with more accurate recognition results.

Till now the propose method is a novel network HomoNet, which can train end-to-end and correct images or features in the LPDR area. The source code of HomoPooling will be open in Github. The following are three main contributions of this paper:

- The proposed unified trainable network is used for fast perspective license plates detection and recognition. By sharing features and building relationships between the two subtasks, the HomoNet not only improves the inference speed but also promotes recognition accuracy in the whole.
- The HomoPooling is introduce which can rectify the distorted license plate or the convolutional features in more complex scenes. The quadrangular region allows it to get more LP information and handle LP from many different perspectives. The sample operation makes it to be differentiable and can be embedded in any network.
- The key point-based detection is apply to replace the bounding box regression. Through the knowledge of one car having one LP, labeling the images by 5 channels of heatmap improves the speed and accuracy of the LP location, as well as removing designed anchors and NMS.

2 Related Works

LPDR is aiming at detecting the position of the license plate and identifying each character in the license plate. Many traditional methods have been proposed respectively [7,9,19]. On account of deep learning (DL) achieves high precision in tasks such as image detection and recognition, researchers have begun to use convolutional networks in LPDR [11,14,15,23]. We will review the DL-based method on license plate detection, text rectification, and end-to-end systems.

2.1 License Plate Detection

The method of license plate detection mainly draws on the model in object detection. YOLO [20] detector is the most popular in this area [8,12,25,30]. Whereas YOLO appears to miss small objects, and YOLOv2 [21] needs artificial designed anchors. Safie et al. [23] used the retinanet algorithm to detect the license plate and solve the problem that the license plate is difficult to detect in scene environments. However, the same as the YOLO, retinanet also needs anchors. All these methods are suitable for multiple targets with different sizes in one image, so they need to design different anchors to match the corresponding target. Recently, researchers have proposed some anchor-free algorithms such as CornerNet [13] and FCOS [27]. CornerNet considered the bounding box as a pair of points and utilized the corner pooling to regress the heat map of each point, which has confirmed to be better than the other single step detector.

2.2 Text Rectification

In the text spotting task, some scholars have proposed different algorithms for correcting the deformed text. Current methods [10,18,26,32] can be classified into affine transformations and non-affine transformations. Spatial Transformer Networks (STN) [10] directly predicted 6 affine coefficients unsupervised trained and rectified input image or feature, adopting a sampler to make it differentiable to end-to-end trainable. Silva et al. [26] proposed the warp-net to regress the affine matrix with supervised information. Liu et al. [18] focused on the rotation of oriented text and proposed ROIRotate to align the text. Although these affine methods could solve some problems to some extent, they could not focus on more useful information. Yang et al. [32] proposed Symmetry-constrained Rectification Network (ScRN) based on Thin-Plate-Spline transformation to solve the curved text, but it was not suitable for rectangular objects.

2.3 End-to-End Systems

There were very few end-to-end networks in the LPDR so that we only found two frameworks. Based on faster-rcnn [22], after the several Convolutional Neural Network (CNN) layers, Li et al. [15] utilized region proposal network to extract the features of interest and applied to detect and identify the license plate at

the same time. Li et al. also designed 6 scales of anchors to match each pixel which limited the inference speed. RPnet [31], based on SSD [17], selected three levels of features to predict the license plate position directly, then used ROI pooling [3] to concatenate the features to identify 7 LP characters which have constant quantities. The above two networks regressed the bounding box while they ignored the deformation of the license plate.

3 Method

HomoNet is an end-to-end trained network that supports license plate detection, rectification and recognition in different natural scenes. There are three submodules: key point detection module, HomoPooling module and recognition module. These three submodules will be

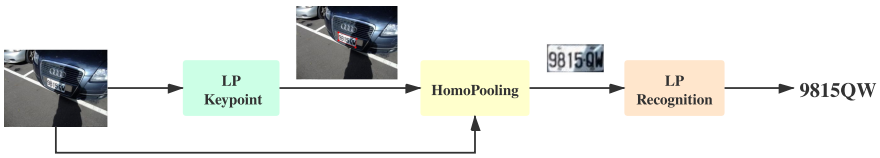

Fig. 2. Pipeline of the proposed method. Three submodules: key point detection module, HomoPooling module and recognition module.

3.1 Overview

The whole framework is shown in Fig. 2. The first is the keypoint detection module. When getting the license plate key points, it can apply them to the perspective transformation to correct the license plate to the horizontal direction, which is easier to recognize. HomoPooling can do the same thing, and it can be end-to-end trained. The last part is the recognition module, which using CNN and max pooling to encode text information, and using Connectionist temporal classification (CTC) [4] to decode the final result.

3.2 LP Keypoint Module

Same as the FOTS [18], LP keypoint module adopts the structure (see Fig. 3) whose backbone is ResNet [5] and Feature Pyramid Network (FPN) proposed in the paper [16]. Firstly, this module concatenates multi-level features from the input image. Only the level of FPN with the stride of 4 is used to extract the feature maps. Then, 4 same conv_gn_relu operations are followed. All convolutions are 3x3 with the stride of 2 and padding of 1. We use group normalization [29] with 32 groups and rule activate function refer to the paper. Finally, one convolution is applied to output 5 channels key points heat map. The first 4 channels

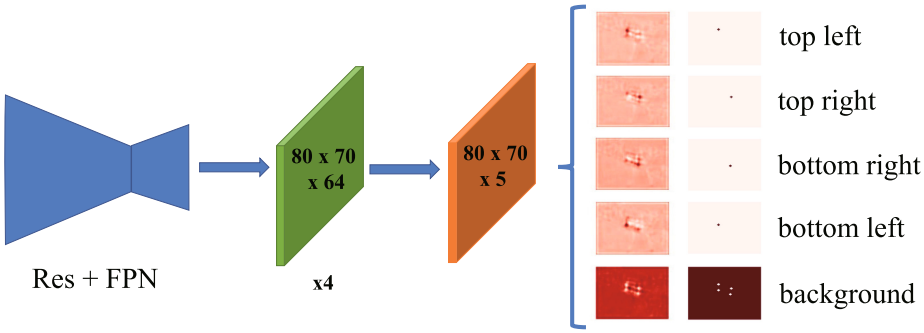

Fig. 3. The LP keypoint module.

predict the probability of every-pixel being the key point, whose order is left-top, right-top, right-bottom and left-bottom. The last channel is background heat map.

The heat map is set to 1 at the position of license corner and 0 at the other position in the first 4 channels, while the label is the opposite on the channel 5. In the experiments, it was found that the key points allow a little offset, which could improve the accuracy and robustness of the model. Thus, a radius of 1 is added at four different directions of ground truth (see Fig. 3). The loss function is binary cross entropy with logits loss, which can be formulated as follows:

$$L_{\text{kp}} = -\frac{1}{N}\sum_{i=1}^{HW}\left[y_i \log \sigma\left(x_i\right) + \left(1 - y_i\right)\log\left(1 - \sigma\left(x_i\right)\right)\right] \tag{1}$$

3.3 HomoPooling

HomoPooling sample the quadrilateral areas determined by four key points to align the distorted license plate to the horizontal direction, as shown in Fig. 2. It consists of three steps. The first step is to get the homography matrix. It needs to normalize the source points and target points between -1 and 1, and then calculate the homography matrix through the four corresponding points, which can be described by the following formulas:

$$Q_S = \begin{bmatrix} -1 + \frac{2x_1}{w} & -1 + \frac{2x_2}{w} & -1 + \frac{2x_3}{w} & -1 + \frac{2x_4}{w} \\ -1 + \frac{2y_1}{h} & -1 + \frac{2y_2}{h} & -1 + \frac{2y_3}{h} & -1 + \frac{2y_4}{h} \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{2}$$

$$Q_T = \begin{bmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{3}$$

$$cQ_S = M_H Q_T \tag{4}$$

Table 1. The architecture of recognition module. The input height is 32 and the width is changeable. An example is given below.

Type	Configuration	Out size
conv_gn_relu	3 × 3, 1 × 1, 1 × 1, 64	32 × 96
max_pooling	2 × 2, 2 × 2, 0 × 0	16 × 48
conv_gn_relu	3 × 3, 1 × 1, 1 × 1, 128	16 × 48
max_pooling	2 × 1, 2 × 1, 0 × 0	8 × 48
conv_gn_relu	3 × 3, 1 × 1, 1 × 1, 256	8 × 48
conv_gn_relu	3 × 3, 1 × 1, 1 × 1, 256	8 × 48
max_pooling	2 × 1, 2 × 1, 0 × 0	4 × 48
conv_gn_relu	3 × 3, 1 × 1, 1 × 1, 512	4 × 48
conv_gn_relu	3 × 3, 1 × 1, 1 × 1, 512	4 × 48
max_pooling	2 × 1, 2 × 1, 0 × 0	2 × 48
conv_gn_relu	2 × 2, 0 × 0, 1 × 1, 512	1 × 47
fc	nc	nc

Where c is a non-zero constant, (x_i, y_i) represents the four ground truth keypoints of the license plate from top-left, top-right, right-bottom to left-bottom. (w, h) is the length of the input images or the feature maps, Q_T is the normalized points of the pooling target. With the relationship between Eq. 2, Eq. 3 and Eq. 5, the homography matrix (M_H) can be found through solving a linear system of equations [2].

The second step is to generate the sample grid. It needs to generate w_H points at equal intervals in the horizontal direction, and h_H points at equal intervals in the vertical direction. Then used the homography matric to calculate the sample grids in source features or images. (w_H, h_H) is the size of HomoPooling.

$$\begin{pmatrix} x_i^g \\ y_j^g \end{pmatrix} = M_H \begin{pmatrix} x_i^t \\ y_j^t \end{pmatrix} \tag{5}$$

Where i and j from -1 to 1, at intervals of w and h relatively.

The final step is to produce the rectified image through sampling from the source grids:

$$T_{ij}^c = \sum_n^h \sum_m^w S_{nm}^c k\left(x_{ij}^s - m; \Phi_x\right) k\left(y_{ij}^s - n; \Phi_y\right) \tag{6}$$

Where T_{ij}^c is the target rectified image value at the position of (i, j) in channel c, and S_{nm}^c is the source image value at the position (n, m) in channel c. h, and w is the height and width of source image. $k()$ is the sampling kernel, and $\Phi_x and \Phi_y$ are its parameters. We use the bilinear interpolation refers to STN [10]. The sample makes it differentiable, so the gradients can be backpropagated from the loss.

3.4 LP Recognition Module

LP recognition module aims at predicting a character sequence from the rectified images transformed by HomoPooling. Because the length of license plate numbers is different in some areas. For example, the new-energy plates have 8 characters while most multiple plates have 7 characters in China, we choose CTC decoder which is different from CCPD [31] and warp-net [26]. Furthermore, due to each character in the license plate is independent, CTC decoder has no connection with context in semantic, which makes us discard the long short term memory (LSTM) module [6] even though it was mostly used in license plate recognition area. What's more, the CTC decoder reduces the computation and speeds the inference.

Inspired by Convolutional Recurrent Neural Network (CRNN) [24], we design recognition module illustrated in Table 1. There are two main operators: conv_gn_relu and max pooling. The configuration means the parameters of convolutional layers and max pooling layers. The parameters are the size of kernel, stride, padding size in height and width as well as channels. nc represents the length of the encoder sequence after a fully-connect layer, which is equal to the number of character dictionary. The "out size" is the feature map size of convolutional layers or sequence length of the encoder layer.

The main differences between our method and CRNN are as follows: First, the group normalization replacing batch normalization which is sensitive to batch choice. Limited by end-to-end training and the memory of GPU devices, training the key points module needs large input size and small batch size. A study [29] indicated that when batch size is small, group normalization performs better than batch normalization. Second, two layers of the LSTM module is removed, which is not suitable for license plate recognition. Third, the input is colorful images or features while the CRNN is gray images.

To preserve the text information in the license plate, conv_gn_relu and max pooling operations are exploited to make the height of the feature maps reduce to 1 and keep the width axis reduce only once. Then, followed the fully-connected layer, we permute the features to time form as a sequence to encode the sequence in a probability distributions over the character dictionary. Finally, CTC is applied to decode the distributions to the recognition results. With the encode sequence $x \in \{x_1, x_2, ..., x_w\}$ and ground truth label $y^* = \{y_1, y_2, ..., y_L\}$, $L \leq w$, the recognition loss can be formulated as:

$$L_{recog} = -\frac{1}{N}\sum_{n=1}^{N} \log p\left(y_n^*|x\right) \tag{7}$$

$$p\left(y^*|x\right) = \sum_{\pi \in B^{-1}(y^*)} p(\pi|x) \tag{8}$$

where $p\left(y^*|x\right)$ is the conditional probability of a given labeling y^* as the sum of the probabilities of all the paths π. B is a many-to-one map from the set of possible labeling groups to the true label. N is the number of input images, and

y_n^* is the corresponding label. For given inputs, we'd like to train our model to maximize the log-likelihood of the summation of Eq. 7.

We trained the keypoint module and recognition module end-to-end, with the following objective function:

$$L = L_{kp} + \lambda L_{recog} \tag{9}$$

Where λ is a hyper-parameter controlling the balance of two losses, which was set to 1 in our experiments.

Fig. 4. Keypoints results of some examples. Images in the first line are from RP, the second line are from CCPD, the third line are from BGY. The images are cut from raw image for a better view.

3.5 Training Skills

HomoNet conducts a pre-trained model on the ImageNet [1] dataset. The Proposed training process involves two stages. In the first stage, it only train the key point module until it has a good performance on the test data. Then HomoPooling is applied to rectify the input image with the ground truth key points and recognition module is added to train jointly. On this basis, we froze the backbone layers and fine-tune it. In the whole training process, there is no any synthetic data or other license plate dataset.

Fig. 5. Rectification LPs from Fig. 4 and recognition results. First line LPs is scale in width for a better view.

4 Experiments

4.1 Datasets

The Application-Oriented License Plate (AOLP), CCPD and BGY datasets were used to verify the proposed model.

AOPL-RP: AOLP has 2046 Chinese Taiwan license plates which has three subsets: Access Control (AC), Traffic Law Enforcement (LE), and Road Patrol (RP). RP owns a wider range of variation in orientation. To compare with the Hsu et al. [9], AC and LE was consider as training set in this article.

CCPD: CCPD, collected from roadside parking in Anhui province of China with 250k images, which is the largest and the most diverse public dataset in LPDR. It has 8 subsets: Base, DB, FN, Rotate, Tilt, Blur, Weather and Challenge. Having great horizontal and vertical degrees, Rotate and Tilt subsets bring difficulty in recognition.

BGY: A new license plate dataset collected from China is introduced in this paper. The dataset contains three license plate colors: blue, yellow and green. There are a total of 2492 pictures. We divide the training set and test set according to 1247 and 1245 respectively. Compared with the public data set, the main features of BGY are: (1) license plates have different background colors; (2) the number of license plates is not fixed; (3) the view, scene, and environment are richer.

Table 2. The configuration of three datasets.

Dataset	Input size	LP input size	Backbone	IoU
RP	240 × 320	32 × 64	Resnet50	0.5
CCPD	480 × 320	32 × 96	Resnet18	0.7/0.6
BGY	512 × 416	32 × 96	Resnet50	0.5

4.2 Implementation Details

We adopt the Stochastic Gradient Descent (SGD) with hyper-parameters (momentum = 0.9, weight decay = 0.0001) to minimize the objective function. Warm up strategy is utilized at the first epoch to increase the learning rate (LR) from 0 to 0.05 in the first stage. Then LR is dropped by 10 times at the 16th epoch and 21st epoch until finishing the training at the 25th epoch. The batch size is set to 8 and the number groups of group normalization is 32. HomoPooling operation resizes the longer side of the input image to times of 32. The proposed model is implemented in PyTorch and trained on Nvidia GTX 1080TI graphics cards.

To improve the robustness of the HomoNet, the experiments involve the data augmentation during training. In the first stage we use the random rotate between $-5°$ to $5°$, and then utilize random disturbance at every key point in the second stage. To be emphasized, there are not any synthetic data.

For different datasets, the input size is changed, the corrected LP size and backbone, as given in Table 2. For a fair comparison with others, we exploit the same metrics for RP and BGY. When the intersection-over-union (IoU) is more than 0.5, the detection is right. When all characters of LP is right and detection is right, the end-to-end result is right. Differently for CCPD, while IoU is more than 0.7, the detection is right. Only when the IoU is more than 0.6 and all characters are right, the end-to-end result is right.

Table 3. Experiments results on the RP subset.

Model	IoU (%)	Speed (ms)	End-to-end (%)	Speed (ms)
Hsu et al. [9]	94	–	85.7	320
Li et al. [14]	95.6	–	88.38	1000
Li et al. [15]	98.2	–	83.63	400
Kessentini et al. [11]	99.2	36.7	90.42	48.7
Silva et al. [26]	–	–	93.29	200
Ours (no rectified)	**99.8**	**18.1**	80.85	**22.5**
Ours	**99.8**	**18.1**	**95.58**	**22.5**

4.3 Performance on RP Dataset

The first line in Fig. 4 shows some of the experimental results trained using our key point module. It can be seen that this module can effectively detect the key points of the license plate even in the dark, strong light, rotation, tilt, natural scene and other conditions. Although some key points have errors, the four key points can still contain the entire license plate information, such as the second in the first line of Fig. 4. At the same time, as illustrated in Table 3, our method achieved the highest performance (99.8%/18.1 ms), 0.6% higher and 18.6 ms faster than the state-of-the-art method (99.2% vs 99.8%, 36.7 ms vs 18.1 ms). The results indicate that Hsu et al. [9] used the expectation-maximization clustering based on the license plate boundary information, and Li et al. [14], regressed box based on the character information can not get high detection accuracy. Both Li et al. [15] who used region proposal network to extract the region of interests and then used anchors to return offset and Kessentini et al. [11] who adopted YOLOv2 detector can get better accuracy. Thanks to the prior information that only one car has only one license plate, we convert the license plate detection into license plate key point detection, which is without anchors and NMS. Our model achieves the best precision and speed.

As shown in Table 3. Extracted local binary patterns (LBP) features and used linear discriminant analysis (LDA) to classify characters [9] can not get a high accuracy. Method [14] which is based on LSTM [6] identifying the character sequence is too slow. The results of the bounding box detection [11,15], which are used to recognize the license plate number, can not get good performance too, while the correction network [26] can improve the recognition. However, affine transformation only can solve a part of the distortion problem. Benefited from rectifying the license plate by HomoPooling with the four points information, HomoNet achieves the highest end-to-end accuracy and the fastest speed. The pooling operation improves a large margin compared to baseline which is directly cropping the image with a bounding box. To further demonstrate the effectiveness of the HomoPooling, Fig. 5 shows the correction results after the HomoPooling operation detected in Fig. 4. It can be seen that the license plates that are perspective, rotated, tilted and distorted are corrected to the horizontal direction.

Table 4. The end-to-end results (percentage) of CCPD and speed is in milliseconds. HC is Holistic-CNN [33]. AP denotes average accuracy.

Model	Speed	AP	Base	DB	FN	Rotate	Tilt	Weather	Challenge
Wang et al. [28] +HC	34.5	58.9	69.7	67.2	69.7	0.1	3.1	52.3	30.9
Ren et al. [22] +HC	28.6	92.8	97.2	94.4	90.9	82.9	87.3	85.5	76.3
Liu et al. [3] +HC	27.8	95.2	98.3	96.6	**95.9**	88.4	91.5	87.3	83.8
Joseph et al. [21] +HC	76.9	93.7	98.1	96	88.2	84.5	88.5	87	80.5
Li et al. [15] +HC	333.3	94.4	97.8	94.8	94.5	87.9	92.1	86.8	81.2
Xu et al. [31] +HC	**16.4**	95.5	98.5	96.9	94.3	90.8	92.5	87.9	85.1
Zhang et al. [33]	–	95.4	98.4	**97**	90.6	92.7	93.5	86.9	84.8
Ours	19	**97.5**	**99.1**	96.9	**95.9**	**97.1**	**98**	**97.5**	**85.9**

4.4 Performance on CCPD Dataset

As shown in Table 4, HomoNet achieves the best results on all subsets except DB. Benefited from HomoPooling, our model achieves a large improvement in Rotate, Tilt and Weather, which is 4.4%, 4.5% and 9.6% respectively. Though the end-to-end speed is 2.6 ms slower than Xu et al. [31] due to the recognition module, our detection module speed is 4 ms faster. The visual results in the second line of Fig. 4 and Fig. 5 indicates the effectiveness of HomoPooling.

4.5 Performance on BGY Dataset

As illustrated in Table 5, we obtain a detection accuracy of 99.52% and an end-to-end recognition accuracy of 91.33% on the BGY dataset. The detection results and the corrected recognition results in the third line of Fig. 4 and Fig. 5 indicate that HomoNet can not only correct but also recognize LPs with multiple types and different character lengths.

Table 5. Experiments results on the BGY dataset.

Model	IoU (%)	Speed (ms)	End-to-end (%)	Speed (ms)
Ours	99.52	24.5	91.33	30.1

5 Conclusions

In this paper, a fast unfied network HomoNet that contains HomoPooling based on perspective transformation was proposed, which can correct oblique license plates. Using the a priori information of one plate for one car, the speed and accuracy of the detection were improved by predicting the heat map of the four points directly. HomoPooling can achieve more information and rectify the oriented LPs. Experiments on public datasets have demonstrated the advantage of the proposed method. In future, the proposed method will extend to the double line license plate to be jointly trained in this framework. Moreover, not all the characters in LP are able to recognize even for people in some situations. We plan to explore the representations of fuzzy LP for vehicle re-identification.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
2. Dubrofsky, E.: Homography estimation. Diplomová práce. Univerzita Britské Kolumbie, Vancouver (2009)
3. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)

4. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 369–376. ACM (2006)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
7. Hou, Y., Qin, X., Zhou, X., Zhou, X., Zhang, T.: License plate character segmentation based on stroke width transform. In: 2015 8th International Congress on Image and Signal Processing (CISP), pp. 954–958. IEEE (2015)
8. Hsu, G.S., Ambikapathi, A., Chung, S.L., Su, C.P.: Robust license plate detection in the wild. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
9. Hsu, G.S., Chen, J.C., Chung, Y.Z.: Application-oriented license plate recognition. IEEE Trans. Veh. Technol. **62**(2), 552–561 (2012)
10. Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
11. Kessentini, Y., Besbes, M.D., Ammar, S., Chabbouh, A.: A two-stage deep neural network for multi-norm license plate detection and recognition. Expert Syst. Appl. **136**, 159–170 (2019)
12. Laroca, R., et al.: A robust real-time automatic license plate recognition based on the YOLO detector. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–10. IEEE (2018)
13. Law, H., Deng, J.: CornerNet: detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
14. Li, H., Shen, C.: Reading car license plates using deep convolutional neural networks and LSTMs. arXiv preprint arXiv:1601.05610 (2016)
15. Li, H., Wang, P., Shen, C.: Toward end-to-end car license plate detection and recognition with deep neural networks. IEEE Trans. Intell. Transp. Syst. **20**(3), 1126–1136 (2018)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
17. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
18. Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J.: FOTS: fast oriented text spotting with a unified network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5676–5685 (2018)
19. Llorca, D.F., et al.: Two-camera based accurate vehicle speed measurement using average speed at a fixed point. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 2533–2538. IEEE (2016)
20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
21. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)

22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
23. Safie, S., Azmi, N.M.A.N., Yusof, R., Yunus, M.R.M., Sayuti, M.F.Z.C., Fai, K.K.: Object localization and detection for real-time automatic license plate detection (ALPR) system using RetinaNet algorithm. In: Bi, Y., Bhatia, R., Kapoor, S. (eds.) IntelliSys 2019. AISC, vol. 1037, pp. 760–768. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-29516-5_57
24. Shi, B., Bai, X., Yao, C.: 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. **39**(11), 2298–2304 (2016)
25. Silva, S.M., Jung, C.R.: Real-time Brazilian license plate detection and recognition using deep convolutional neural networks. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 55–62. IEEE (2017)
26. Silva, S.M., Jung, C.R.: License plate detection and recognition in unconstrained scenarios. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 593–609. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_36
27. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355 (2019)
28. Wang, S.Z., Lee, H.J.: A cascade framework for a real-time statistical plate recognition system. IEEE Trans. Inf. Forensics Secur. **2**(2), 267–282 (2007)
29. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
30. Xie, L., Ahmad, T., Jin, L., Liu, Y., Zhang, S.: A new CNN-based method for multi-directional car license plate detection. IEEE Trans. Intell. Transp. Syst. **19**(2), 507–517 (2018)
31. Xu, Z., et al.: Towards end-to-end license plate detection and recognition: a large dataset and baseline. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 255–271 (2018)
32. Yang, M., et al.: Symmetry-constrained rectification network for scene text recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9147–9156 (2019)
33. Zhang, Y., Huang, C.: A robust Chinese license plate detection and recognition system in natural scenes. In: 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), pp. 137–142. IEEE (2019)