Tripartite Architecture License Plate Recognition Based on Transformer

Ran Xia[1], Wei Song[1,2,3](✉), Xiangchun Liu[1], and Xiaobing Zhao[1,2,3]

[1] School of Information Engineering, Minzu University of China, Beijing 100081, China
songwei@muc.edu.cn

[2] National Language Resource Monitoring and Research Center of Minority Languages, Minzu University of China, Beijing 100081, China

[3] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China

Abstract. Under natural conditions, license plate recognition is easily affected by factors such as lighting and shooting angles. Given the diverse types of Chinese license plates and the intricate structure of Chinese characters compared to Latin characters, accurate recognition of Chinese license plates poses a significant challenge. To address this issue, we introduce a novel Chinese License Plate Transformer (CLPT). In CLPT, license plate images pass through a Transformer encoder, and the resulting Tokens are divided into four categories via an Auto Token Classify (ATC) mechanism. These categories include province, main, suffix, and noise. The first three categories serve to predict the respective parts of the license plate - the province, main body, and suffix. In our tests, we employed YOLOv8-pose as the license plate detector, which excels in detecting both bounding boxes and key points, aiding in the correction of perspective transformation in distorted license plates. Experimental results on the CCPD, CLPD, and CBLPRD datasets demonstrate the superior performance of our method in recognizing both single-row and double-row license plates. We achieved an accuracy rate of 99.6%, 99.5%, and 89.3% on the CCPD Tilt, Rotate, and Challenge subsets, respectively. In addition, our method attained an accuracy of 87.7% in the CLPD and 99.9% in the CBLPRD, maintaining an impressive 99.5% accuracy even for yellow double-row license plates in the CBLPRD.

Keywords: License Plate Recognization · License Plate Detection · Transformer

1 Introduction

Automatic License Plate Recognition (ALPR) is a computational system that automatically detects and recognizes license plates from images or videos using computer vision and machine learning technologies. Compared to pure Latin character license plates, Chinese license plate recognition proposes additional

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14426, pp. 412–423, 2024.
https://doi.org/10.1007/978-981-99-8432-9_33

challenges. Chinese license plates have two structures: single line and double line, and the algorithm's adaptability to non-single-line license plates needs to be further considered. Chinese license plates include characters that represent 34 provincial-level administrative regions, which increases the types of characters to be recognized, and certain license plates feature distinct suffixes, such as '挂' , '学' , '警' , etc. Compared to Latin characters, Chinese characters have a high degree of glyph complexity and similarity, making recognition more challenging. These characters are also more susceptible to misidentification due to factors such as lighting conditions, blur, and shooting angles.

In this paper, we propose a novel Chinese License Plate Transformer (CLPT) for the recognition of Chinese license plates. This system is inspired by the transformer model's [11] remarkable performance and adaptability in various vision tasks. Our method has the following four insights:

1. We propose a Tripartite Architecture(TA) that deconstructs all Chinese license plates into a province-main-suffix format. This separation not only comprehensively encompasses all types of Chinese license plates but also enables the model to leverage the inherent structural characteristics of license plates. Consequently, this approach significantly enhances the accuracy of license plate recognition.
2. We propose an Auto Token Classify(ATC) mechanism, designed to complement the TA architecture. This mechanism adaptively categorizes all output tokens from the transformer into several groups, aligning with specific subtasks, including province classification, recognition of the Latin character main body, and suffix classification.
3. Compared to conventional algorithms that directly utilize YOLO for license plate detection, the extension of YOLOv8, known as YOLOv8-pose, offers the capability to predict the four key points of a license plate additionally, without imposing a significant computational burden. Harnessing these key points for perspective transformation correction enhances the model's proficiency in recognizing license plates under distorted viewing angles.
4. Compared to traditional Recurrent Neural Networks (RNNs), the utilization of a Transformer architecture does not inherently constrain recognition content to a predefined left-to-right single-line sequence. As a consequence, Transformer models demonstrate notable advantages in the recognition of dual-line license plates.

2 Related Work

2.1 License Plate Recognition

Raj et al. [7] segmented the characters on the license plate for OCR recognition. However, this method is dependent on the segmentation model and therefore has an error accumulation issue. Xu et al. [15] proposed RPnet, an end-to-end license plate recognition system that finally uses seven classifiers to predict the characters on the license plate separately. However, this method can only identify

7-digit blue plates rigidly. GONG Y et al. [3] proposed predicting the rotation angle θ in the LPD part to correct the rotated license plate and then use CTC to predict characters of variable length. However, this method struggles with non-planar rotated license plates; Wang et al. [13] proposed a shared-parameter classification head for the CCPD dataset, which segments the prediction of blue license plates into province, city alphabet, and a sequence of five Latin characters. However, this method presents challenges when attempting to apply it universally to other types of license plates.

2.2 Transformer

Transformers have shown good performance in the field of natural language processing. ViT [2] (Vision Transformer) applied Transformers to the visual field and achieved excellent results. However, due to the large architecture and slow inference speed of ViT, it is limited in its application in the license plate recognition task. With the birth of lightweight Transformers, such as MobileViT [6], Deit [10], etc., we propose a new solution to this problem. Wu et al. [14] proposed TinyViT, a new family of tiny and efficient vision transformers, pretrained on large-scale datasets with their proposed fast distillation framework. While ensuring lightweight and high efficiency, TinyViT possesses a hierarchical structure that can better handle the detailed features in Chinese characters. We use the lightest TinyViT-5M as the pre-training encoder, divide the output results into three sub-tasks of province and suffix classification, and Latin character body sequence recognition through the ATC mechanism. This not only achieves excellent license plate recognition performance, but also provides a new way of thinking for using Transformer models for license plate recognition.

3 Proposed Method

3.1 License Plate Detection

Specifically in license plate detection tasks, the bbox-based YOLO algorithm may face accuracy issues due to possible rotation and distortion of the license plate, as its rectangular representation struggles to capture the detailed characteristics of these distortions.

We adopted the YOLO-Pose algorithm, an extension of the traditional YOLO, which includes key point prediction. This feature provides a significant advantage over regular YOLO algorithms in addressing rotated license plates. By modifying the detection head, our model simultaneously predicts the bbox and the four corner points of the license plate. This key point information enables us to perform a perspective transformation, effectively correcting for rotation and distortion without significantly adding complexity to our approach.

3.2 License Plate Recognition

Tripartite Architecture(TA). As shown in Fig. 1, our Tripartite Architecture (TA) partitions the license plate into three components: province, main body, and suffix. This strategy, which leverages the structural information of the license plate, positions the province and suffix at fixed points, each containing specific Chinese characters.

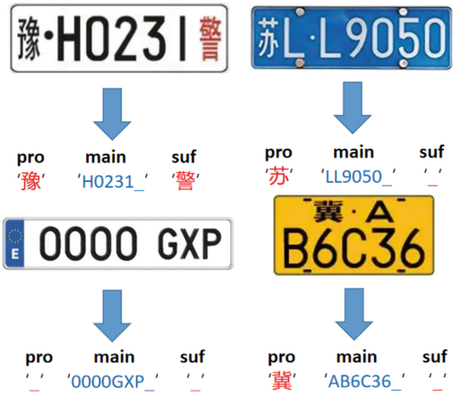

Fig. 1. partitioning of Different Types of License Plates

This design allows the main body to focus solely on Latin characters, reducing the prediction task complexity. To further enhance flexibility for various license plate structures, we incorporate the ‘_’ character. Within the province and suffix, it represents an absent character, while it signals sequence termination in the main body. This design handles different license plate structures and lengths effectively, showcasing strong generalization capabilities.

CLPT. As shown in Fig. 2, the network mainly consists of a Transformer encoder with a pyramid structure, an ATC module, and post-processing corresponding to the token. The 224×224 image is first encoded into a series of tokens that enter the Transformer Encoder after the Patch Embedding process. The Transformer Encoder of TinyViT consists of downsampling three times and Transformer Block, forming a hierarchical structure. Specifically, downsampling in the Transformer Encoder uses MBConv, and the Swin structure is used in the Transformer Block to perform self-attention on tokens within the window. This process gradually downsamples the original encoded 56×56 tokens to 7×7 tokens. After the encoding is complete, each of the 49 tokens of 320 dimensions contains features of the corresponding patch after sufficient self-attention interaction. At this time, the ATC module performs soft grouping on these 49 tokens. For noise tokens, we do not do any subsequent processing; for province and suffix tokens, we perform a global average pooling on these tokens, followed by a fully connected layer, to classify the province character or suffix character. For the variable-length Latin character sequence, we need to select n key tokens from it for sequence prediction. n represents the maximum length of the middle character sequence, which includes a special symbol ‘__’ to indicate the end of the

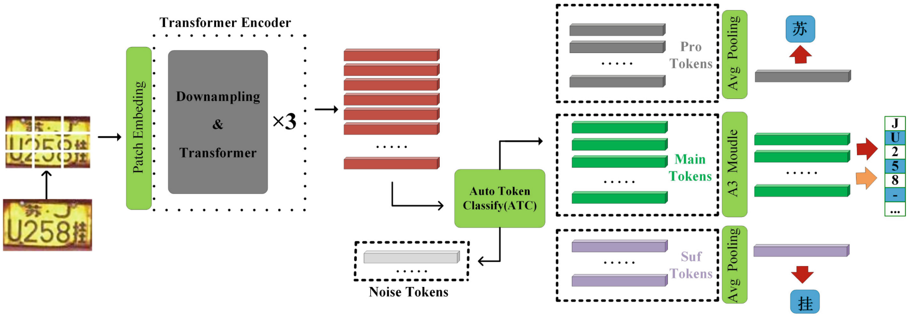

Fig. 2. Flowchart of CLPT

sequence. We use the A3 (Adaptive Addressing and Aggregation) module proposed by Wang et al. [12] to adaptively weight the results through the spatial attention mechanism and fuse n key tokens containing character information from these 49 tokens. We connect a fully connected layer after key tokens to get the prediction of the main part. All three prediction processes use cross-entropy as the loss function, and the sum is added in a certain ratio to obtain the final loss function formula as follows:

$$\text{Loss} = \lambda_1 \cdot \text{Loss}_{\text{pro}} + \lambda_2 \cdot \text{Loss}_{\text{main}} + \lambda_3 \cdot \text{Loss}_{\text{suf}} \tag{1}$$

where λ is the weight of different losses. In our method, λ_1, λ_2, and λ_3 are set to 0.35, 0.5, and 0.15, respectively.

Auto Token Classify(ATC) Mechanism. Once the image has passed through the encoder, each token obtains its corresponding information. We use the ATC mechanism to softly classify the output tokens, allowing tokens containing specific corresponding information to complete different tasks. Specifically, some tokens contain provincial information, some contain information of the Latin character main body, and some contain suffix information. In addition, we added a Noise Token category to store tokens that primarily contain noise (Fig. 3).

For a token classifier with four categories (province, main body, suffix, and noise), we first map the input x to the scores of the four categories. This can be expressed as:

$$\mathbf{s}(x) = \mathbf{W}_2\text{ReLU}(\text{LayerNorm}(\mathbf{W}_1 x + \mathbf{b}_1)) + \mathbf{b}_2 \tag{2}$$

where $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, and $\mathbf{b}_2$ are the weight and bias parameters of the network.

Then, we use the softmax function to transform the output of the classifier into a probability distribution:

$$\mathbf{p}(x) = \text{softmax}(\mathbf{s}(x)) \tag{3}$$

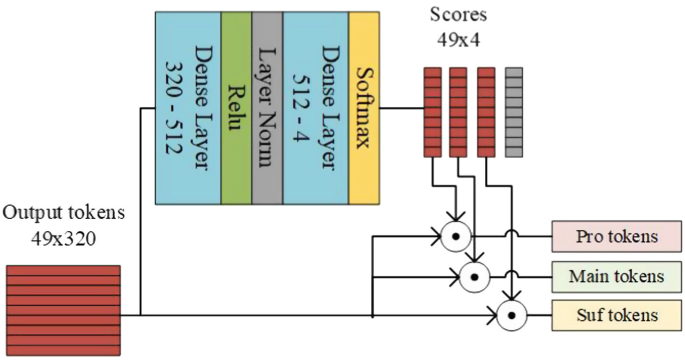

Fig. 3. ⊙ represents the element-wise multiplication operation. The gray part in the "scores" indicates the proportion of Noise Tokens (Color figure online)

Finally, the ATC computes the result tensor corresponding to each category, i.e., province, Latin character main body, and suffix, excluding the noise token. This process involves multiplying the input vector element-wise by the probability of each category:

$$\mathbf{r}_i = x \odot \mathbf{p}_i(x) \quad \text{for } i \in 0, 1, 2 \tag{4}$$

where ⊙ represents the element-wise multiplication operation, and $\mathbf{p}_i(x)$ is the output of the softmax function, corresponding to the probability of category i. Finally, we obtain the result tensor $\mathbf{r}_i$ corresponding to each category (province, main body, suffix), which encodes the information related to each category in the input vector x.

4 Experiments

4.1 Datasets

CCPD Dataset. The Chinese City Parking Dataset (CCPD) [15] is a large license plate recognition dataset comprising about 290k images from various parking lots in China. The dataset includes the following subsets: CCPD-base (200k), CCPD-db (20k), CCPD-fn (20k), CCPD-rotate (10k), CCPD-tilt (10k), CCPD-weather (10k), CCPD-challenge (10k). Half of the CCPD-base subset is used for training, while the remaining subsets are utilized for testing.

CLPD Dataset. The Comprehensive License Plate Dataset (CLPD) is a richly-annotated dataset containing 1200 images of various types of vehicles, covering all 31 provinces in mainland China, with diverse shooting conditions and regional codes. Notably, the dataset includes both seven-letter and eight-letter license plates, presenting increased recognition complexity, and making it a significant tool for our experimental setup.

CBLPRD Dataset. The “China-Balanced-License-Plate-Recognition-Dataset-330k” (CBLPRD) is an open-source, large-scale dataset containing 330k Chinese license plate images produced by Generative Adversarial Networks (GANs). The images in this dataset are of excellent quality and cover a variety of Chinese license plate types.The dataset consists of 300,000 training images and 10,000 validation images, supporting the training and validation of models. In particular, it includes some license plate types that are rare in other datasets, such as yellow double-row license plates, embassy license plates, and tractor green plates, which add to the value and importance of the dataset.

4.2 Experimental Environment and Tools

Our network was run on a computer with a 24G RTX 3090 graphics card and an 11th generation Intel Core i7-11700K processor. We implemented the deep learning algorithm based on Pytorch. For yolov8-pose, we used the Adam optimizer, set the batch size to 128, set the learning rate to 0.01, and used mosaic enhancement and random perspective transformation. For CLPT, we used the Adadelta optimizer, set the batch size to 128, set the learning rate to 1, and did not use any data augmentation.

5 Results

5.1 CCPD

For license plate detection, we utilized the method proposed by [15], focusing solely on precision. A prediction is deemed correct if the Intersection over Union (IoU) between the predicted bounding box and the ground truth exceeds 0.7. As presented in Table 1, YOLOv8 outperforms the other methods across all subsets, particularly achieving a 7.0% and 5.8% boost in the Rotate and Challenge subsets, respectively.

Table 1. Comparison of the average precision (percentage) of license plate detection in different subsets. AP represents the average accuracy of the entire dataset.

Method	AP	Base	DB	FN	Rotate	Tilt	Weather	Challenge
Faster-RCNN [9]	92.9	98.1	92.1	83.7	91.8	89.4	81.8	83.9
TE2E [4]	94.2	98.5	91.7	83.8	95.1	94.5	83.6	93.1
RPnet [15]	94.5	**99.3**	89.5	85.3	94.7	93.2	84.1	92.8
YOLOv4 [1]	95.1	96.8	93.7	93.1	93.5	94.7	96.6	85.5
YOLOv3 [8]	96.0	97.1	97.2	93.3	91.6	94.6	97.9	90.5
YOLOv8	**99.0**	**99.3**	**99.1**	**98.8**	**98.6**	**99.2**	**99.7**	**96.3**

For combined license plate detection and recognition, a positive sample is confirmed when the IoU between the bounding box and ground truth surpasses 0.6 and all characters are predicted correctly. We tested using both bounding box results (without rotation correction) and keypoint results (with distortion correction). As depicted in Table 2, aside from the DB subset, our method yielded the best results.

Table 2. The table compares license plate detection and recognition accuracy across various subsets, distinguishing between using bounding boxes ([bbox]) and corrected keypoints ([keypoint]), both detected by YOLOv8.

Methods	Base	DB	FN	Rotate	Tilt	Weather	Challenge
Ren et al. [9]	92.8	97.2	94.4	90.9	82.9	87.3	76.3
Liu et al. [5]	95.2	98.3	96.6	95.9	88.4	91.5	83.8
Xu et al. [15]	95.5	98.5	96.9	94.3	90.8	92.5	85.1
Zhang et al. . [16]	93.0	99.1	96.3	97.3	95.1	96.4	83.2
Zhou et al. [17]	97.5	**99.2**	98.1	98.5	90.3	95.2	86.2
ours[bbox]	**99.8**	**99.2**	**98.8**	98.5	98.8	**98.3**	**89.3**
ours[kepoint]	**99.8**	98.9	**98.8**	**99.5**	**99.6**	98.1	89.0

Keypoint correction notably enhances the accuracy in handling rotated license plates (Rotate and Tilt subsets) improving the results by 1.6% and 0.2% respectively, however, it shows a slight decrease of 0.1% on the FN and Weather subsets, and 0.5% on the DB subset. Conversely, predicting with bounding boxes results in higher accuracy when the license plates are brighter or darker (DB subset) (Fig. 4).

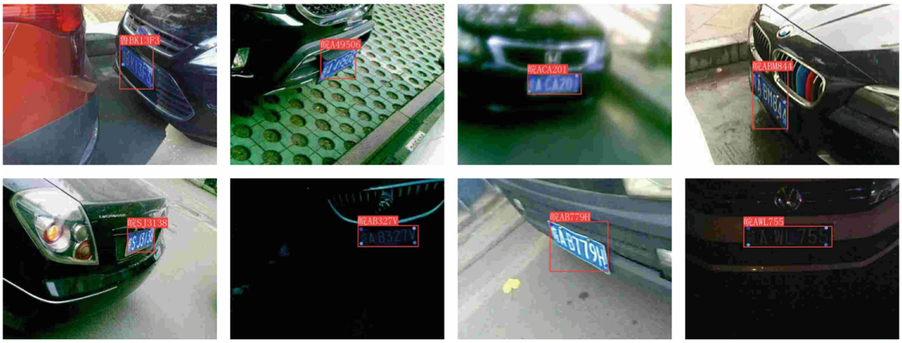

Fig. 4. Results display in CCPD dataset

5.2 CLPD

In the context of automatic license plate recognition, the generalization capability of a model holds significant importance as an evaluation metric. Following the methodology proposed in reference [16], we exclusively employ the base subset of the CCPD dataset for training, while utilizing the CLPD dataset, which encompasses license plate samples from various diverse scenarios, as our test set. By adopting this approach, the results obtained from the CLPD dataset effectively showcase the model's ability to generalize to other datasets. This evaluation technique offers a comprehensive validation of the model's performance across different scenarios and conditions. We consider only those license plates as positive samples which have been entirely and correctly identified, given the fact that only completely accurate predictions bear practical meaning in license plate recognition. The experimental results on the CLPD dataset are presented in Table 3. With the base subset of CCPD as our training set, without any addition of synthetic license plates, our method achieves a Top1 accuracy of 83.4% on the CLPD. Further, with the inclusion of the CBLPRD dataset as additional augmentation data, we manage to reach a Top1 accuracy of 87.7%.

Table 3. Comparison of License Plate Recognition Accuracy on CLPD Dataset

Method	Accuracy
Xu et al. [15]	66.5
Zhang et al. (real data only) [16]	70.8
Zhang et al. (real + synthetic data) [16]	76.8
Zou et al. [17]	78.7
ours(real data only)	**83.4**
ours(real+synthetic data)	**87.7**

5.3 CBLPRD

RNNs carry an inherent assumption that characters are arranged in a sequence from left to right and are contained in a single line. For instance, the classical Convolutional Recurrent Neural Network (CRNN) compresses the original image height to one during the CNN process, thereby inputting it into the subsequent RNN. Although LPRNet does not utilize an RNN, it primarily extracts horizontal features using a 1×13 convolutional kernel and then transforms the feature map into a sequential format via height-wise pooling. These methods show limitations when faced with double-row license plates. A bidirectional RNN can alleviate this issue to some extent.

However, in contrast, our proposed Transformer-based model is not limited by these assumptions. Its feature vectors, extracted by the adaptive addressing

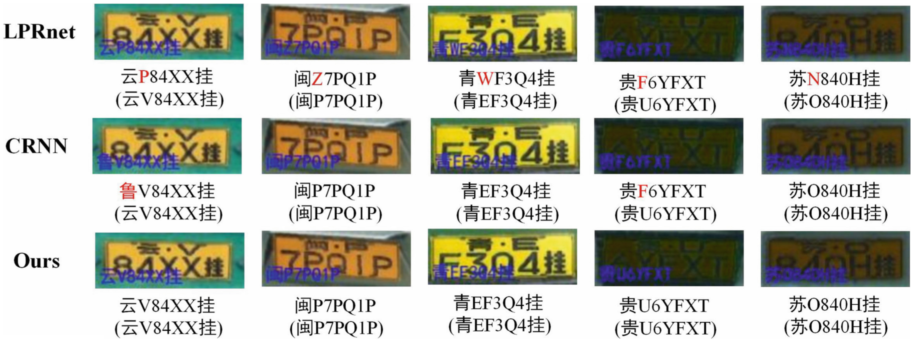

Fig. 5. The figure shows the recognition results of different algorithms on two-way license plates. The words in parentheses are the ground truth.

module, can adapt more effectively to various license plate structures. In Fig. 5, we observe that for LPRNet and CRNN, all misclassifications occur in the first row of characters in dual-row license plates. This underscores the difficulty these algorithms face in accurately identifying license plates with dual-row structures. In contrast, our proposed method demonstrates commendable proficiency in the precise recognition of such dual-row license plates. We conducted experiments on the CBLPRD dataset and, in addition, listed and tested the accuracy of yellow double-row license plates in the validation set to observe the algorithm's ability to recognize double-row license plates. The experimental results shown in Table 4 confirmed our hypothesis: on the validation set, the accuracy of LPRNet is 84.3%, but it cannot recognize double-row license plates; the bidirectional RNN (BiLSTM) in CRNN alleviates this problem to a certain extent, but when processing double-row license plates, the accuracy still significantly decreases by 8.9%. Our Transformer model has an accuracy of 99.9% on the validation set, and when dealing with yellow double-row license plates, the accuracy remains as high as 99.5%.

Table 4. Performance Comparison of License Plate Recognition Algorithms on Validation Set, with Additional Emphasis on Yellow Double-row Plates Accuracy

Algorithm	Validation Set	Yellow Double-row Plates
LPRNet	84.3	0.0 (−84.3)
CRNN	97.7	88.8(−8.9)
CLPT(ours)	**99.9**	**99.5** (−0.4)

These experimental results clearly show that, compared to RNN-based models, our Transformer model demonstrates exceptionally high adaptability and robustness when recognizing complex license plate formats (such as double-row license plates).

6 Conclusion

This paper introduces a transformer-based license plate recognition framework named Chinese License Plate Transformer (CLPT).By leveraging a Tripartite Architecture(TA) and the ATC mechanism, CLPT effectively manages the complexities inherent to Chinese characters and the distinct structures of Chinese license plates. Furthermore, we have showcased the superiority of YOLOv8 in license plate detection and suggested an extension of YOLOv8, named YOLOv8-Pose. This extension enhances the detection performance for distorted and rotated license plates without imposing a significant additional computational burden.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOV4: optimal speed and accuracy of object detection. arXiv preprint: arXiv:2004.10934 (2020)
2. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint: arXiv:2010.11929 (2020)
3. Gong, Y., et al.: Unified Chinese license plate detection and recognition with high efficiency. J. Vis. Commun. Image Represent. **86**, 103541 (2022)
4. 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)
5. 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
6. Mehta, S., Rastegari, M.: MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint: arXiv:2110.02178 (2021)
7. Raj, S., Gupta, Y., Malhotra, R.: License plate recognition system using yolov5 and CNN. In: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 372–377. IEEE (2022)
8. Redmon, J., Farhadi, A.: YOLOV3: an incremental improvement. arXiv preprint: arXiv:1804.02767 (2018)
9. 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, vol. 28 (2015)
10. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
11. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
12. Wang, P., Da, C., Yao, C.: Multi-granularity prediction for scene text recognition. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022. Lecture Notes in Computer Science, vol. 13688, pp. 339–355. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19815-1_20
13. Wang, Y., Bian, Z.P., Zhou, Y., Chau, L.P.: Rethinking and designing a high-performing automatic license plate recognition approach. IEEE Trans. Intell. Transp. Syst. **23**(7), 8868–8880 (2021)

14. Wu, K., et al.: TinyViT: fast pretraining distillation for small vision transformers. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022. Lecture Notes in Computer Science, vol. 13681, pp. 68–85. Springer, Cham (2022)
15. Xu, Z., et al.: Towards end-to-end license plate detection and recognition: a large dataset and baseline. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 261–277. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_16
16. Zhang, L., Wang, P., Li, H., Li, Z., Shen, C., Zhang, Y.: A robust attentional framework for license plate recognition in the wild. IEEE Trans. Intell. Transp. Syst. **22**(11), 6967–6976 (2020)
17. Zou, Y., et al.: License plate detection and recognition based on YOLOV3 and ILPRNET. SIViP **16**(2), 473–480 (2022)