Sequence-based Recognition of License Plates with Severe Out-of-Distribution Degradations*

Denise Moussa[0000−0002−1390−9198], Anatol Maier[0000−0002−8093−7252], Franziska Schirrmacher[0000−0003−1511−7669], and Christian Riess[0000−0002−5556−5338]

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
{denise.moussa}@fau.de

Abstract. Criminal investigations regularly involve the deciphering of license plates (LPs) of vehicles. Unfortunately, the image or video source material typically stems from uncontrolled sources, and may be subject to severe degradations such as extremely low resolution, strong compression, low contrast or over- resp. underexposure. While LP recognition has a long history in computer vision research, the deciphering under such severe degradations is still an open issue. Moreover, since the data source is not controlled, it cannot be assumed that the exact form of degradation is covered in the training set.

In this work, we propose using convolutional recurrent neural networks (CRNN) for the recognition of LPs from images with strong unseen degradations. The CRNN clearly outperforms an existing conventional CNN in this scenario. It also provides an additional particular advantage for criminal investigations, namely to create top-n sequence predictions. Even a low number of top-n candidates improves the recognition performance considerably.

Keywords: License Plate Recognition · Forensics · Low-quality Images.

1 Introduction

Criminal investigations often include the examination of photo and video recordings that may serve as clue or evidence with probative value in a legal context. Forensic material might also contain the depiction of an escape vehicle or parts thereof. An important task then is the robust detection and recognition of the vehicle's license plate (LP). Unfortunately, such material regularly stems from sources of low quality, for example from cheap surveillance camera systems. In addition, environment based deterioration like motion blur or advert lighting can be challenging.

Most existing works on LP recognition assume high-quality images, i.e. photos that can also be deciphered by humans. Only few works address the task of

* We gratefully acknowledge support by the German Federal Ministry of Education and Research (BMBF) under Grant No. 13N15319

recognizing severely degraded LPs [2, 7, 11, 14]. These works use standard CNN architectures. However, research on high-quality LPs showed improved performance by conducting a sequence analysis with the help of combined CNN and RNN architectures [10, 15, 16, 18, 20]. These types of networks exploit that LP character sequences usually follow rules which can be grasped by such architectures.

In this work, we propose such a sequence analysis for severely degraded LPs that are not perceptible to the human eye. Our adapted CRNN provides two essential benefits for criminal investigations. First, it performs very well on out-of-distribution degradations that were not part of the training set, which is a vital precondition for analyzing footage from unknown sources. Second, it provides top-n predictions, which boosts the detection already for low n. Our specific contributions thus are:

1. We adapt the CRNN architecture by Shi *et al.* [15] for recognizing severely degraded LPs.
2. We show the effectiveness of the proposed method on real data, and evaluate its robustness on out-of-distribution samples on synthetic data.
3. We show that the proposed method outperforms existing work. Major additional improvements are obtained from the CRNN's top-n predictions.

In Sec. 2, we describe previous approaches to LP recognition, particularly on low quality data. Section 3 presents the proposed architecture and data. In Sec. 4, we present the experimental protocol and results. Section 5 discusses our findings, and Sec. 6 concludes this work.

2 Related Work

Automatic license plate recognition (ALPR) is an active research topic, where most works focus on real-world image and video data of good quality that are captured from known cameras. These methods typically utilize a two stage process, consisting of LP detection and character recognition. Detection is often based on the YOLO real-time object detector [1,8,9,12,13,19]. Character recognition is modeled either as a classification task [1,8,9,19], or as a sequence labeling task via convolutional recurrent neural networks (CRNNs) [10, 13, 15, 16, 18, 20]. Here, convolutional layers extract features from an image which is then reshaped into a sequence as input for recurrent layers. The resulting output matrix is then transformed to the final character sequence by a transcription layer using connectionist temporal classification (CTC) [5]. Contrary to CNNs, CRNNs are not limited to individual features for character predictions, but instead they operate on the whole sequence. Furthermore, CRNNs can predict sequences of arbitrary length without changes to the architecture.

However, ALPR on images that are degraded beyond human recognition has only received little attention so far. Agarwal *et al.* [2] propose a CNN to process LPs degraded by noise and very low resolution. Their method classifies two groups of three characters under several constraints on font style, character ratio,

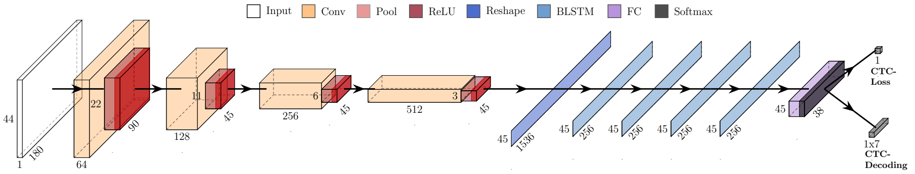


Fig. 1: Configuration of our adapted CRNN architecture

character width and foreground-background contrast. Lorch *et al.* [11] generalize their approach to LPs of five to seven characters without constraints on character properties. Their model uses seven output layers to achieve a per-character prediction for the whole input image. Kaiser *et al.* [7] analyze the impact of lossy compression on the performance of Lorch *et al.* [11]. The authors also study detection accuracy w.r.t. character position and the confusion of characters with similar shape. Recently, Rossi *et al.* [14] investigated cues for interpreting and explaining predictions for severely degraded LPs by adding a U-NET based denoiser prior to the recognition CNN. Hereby, the denoised output allows for a validation of the predictions.

All these previous works model degraded LP recognition by classifying each character independently [2, 7, 11, 14]. However, as for ALPR on high-quality images, we argue that exploiting sequence information also improves recognition on strongly degraded LPs. We propose a CRNN architecture and compare its performance to the baseline method by Lorch *et al.* [11]. Our training data consists of synthetic Czech LPs subjected to compression, low resolution and Gaussian noise as employed by Kaiser *et al.* [7]. We provide an extensive robustness evaluation on degradations that are not part of the training set, namely underexposure, shot noise, salt and pepper noise, motion and defocus blur. We evaluate on real Czech LP images, showing the performance on mixed unseen degradations as expected in practice.

3 Methods

Here, we describe the proposed CRNN model, the pipeline for generating the synthetic training data and all degradation types for our training and test sets.

3.1 CRNN Architecture

Our CRNN model is based on the architecture by Shi *et al.* [15] which consists of three parts: a convolutional part for feature extraction, a recurrent part for sequence prediction, and a CTC layer for loss computation and sequence decoding. We adapt all hyperparameter values to fit our task formulation as described in the following text.

The architecture is shown in Fig. 1. The receptive field of the input is $44 \times 180 \times 1$ pixels. Each sample is first processed by four convolutional blocks with 64, 128, 256, and 512 trainable filter kernels of size 3×3 with stride $z = 1$. Zero-padding is used to avoid shrinking the output. Each convolutional layer is followed by a max pooling layer with stride $z = 1$, where the filter dimensions per layer with increasing depth are $2 \times 2, 2 \times 2, 2 \times 1$, and 2×1. All convolutional blocks conclude with a ReLU activation followed by batch normalization.

The output of the convolutional part is reshaped as described by Shi *et al.* [15] to form a sequence of feature vectors along the columns of the input image. This enables the following bidirectional long short-term memory (BLSTM) [6] layers to sequentially process the feature map column by column. The four BLSTM layers are of depth 256, consisting of one forward and one backward LSTM of depth 128. Each layer uses dropout with probability 0.5. The resulting label sequence is $\mathbf{y} = [\mathbf{y}_0, \ldots, \mathbf{y}_T]$ with $T = 44$, where the 45 image columns are interpreted as positional dimensions in the sequence. This sequence is processed by a 45×37 fully connected layer, where the 37 columns encode the label characters, i.e., the 26 latin characters, the digits 0-9, and a blank. A softmax layer then yields a probability distribution over all 37 labels for each time step t. During training, the CTC loss is computed from the softmax output. During inference, CTC decoding is applied and the resulting prediction is returned.

The convolutional layers and BLSTM layers use Glorot (Xavier) uniform initialization for the kernel weights, and the bias is initialized with zeros. We apply the Adam optimizer with learning rate $\eta = 1\text{e}{-4}$ and default parameters $\mu = 0.9$ and $\rho = 0.999$. Training and evaluation both use a batch size of 32. The training and validation data consists of 10 000 000 and 5 000 samples, respectively. Training is performed for one epoch, so that the model only processes each individual sample once. For the evaluation, we use test sets of 10 000 samples per run.

3.2 Generation of Synthetic Data Sets

To the best of our knowledge, no large-scale real world data exists with controlled degradations. Hence, we train the network on synthetic data. This allows to control degradation types and strengths for analyzing the impact of individual degradations. We generate Czech LPs to also evaluate our method on the real-world data set by Špaňhel *et al.* [17] (cf. Sec. 4.2).

We adapt the data generator by Lorch *et al.* [11] to Czech LP specifications, including the character ratio, character number, font type, different offsets and gap sizes [4]. Each LP consists of seven characters c_n, $0 \leq n \leq 6$, in groups of three and four characters separated by a gap with a sticker. For the first letter, we slightly deviate from the specification and only generate digits, since letters are very rare in practice. Position $c_1 \in \{A, B, C, E, H, J, K, L, M, P, S, T, U, Z\}$ specifies the region. Position c_2 can be any Latin character or digit. Positions $c_3, \ldots, c_6$ can only contain digits.

The pipeline for generating gray scale LPs is implemented as follows. First, random syntactically correct LP character strings are created with fonts, gaps and offsets within the specification range. Then, a background with random speckle is added to support the model generalization to real data. The font intensity is randomly selected while ensuring a minimum contrast to the background. The resulting image has a resolution of 120×520 pixels. Task-specific degradations are applied prior to nearest neighbor downsampling to a size of 44×180 pixels [11]. For network training we scale the intensities to $[0, 1]$.

3.3 Training Set Degradations

During training we use the union of the three types of image degradations from Lorch *et al.* [11] and Kaiser *et al.* [7]. These degradations include downsampling to extremely low resolution to simulate vehicles at a distance, the addition of additive Gaussian noise as general distortion, and the addition of JPEG compression. Note that for practical criminal investigations, further augmentations could be used. However, they generally have to operate on out-of-distribution data, where pictures stem from uncontrolled sources and suffer from all sorts and combinations of degradations. Hence, we argue that this selection approximates a minimum set of degradations that can be anticipated already at training time, while leaving sufficiently large room for unseen degradations to evaluate the method performance.

To each of the 10 000 000 training and 5 000 validation samples, each degradation is applied with a randomly chosen strength. Here, we use bicubic downsampling to create LPs with widths $w \in$ [20,180] pixels, additive Gaussian Noise with a signal-to-noise ratio between original and distorted image of $SNR_{db} \in [-3, 20]$, and JPEG compression with quality factors $q \in \{5, \ldots, 95\}$.

3.4 Test Set Degradations

Real-world LP images from uncontrolled sources may exhibit degradations that are not covered by the training data. We hence create out-of-distribution degradations only used during testing.

Underexposure occurs at low-light acquisitions, which are of particular interest for criminal investigators. We simulate underexposure by multiplying a factor $c < 1$ to every pixel of an image.

Shot Noise is created within the camera, and has larger impact on low-light acquisitions. In the simulation, we synthetically underexpose the image and draw each final pixel value from a Poisson distribution with standard deviation λ equal to the underexposed pixel value [3].

Salt and Pepper Noise is an impulse noise, randomly appearing as dark and bright pixels, typically caused by technical defects. We simulate this noise by randomly setting pixel values to 1 (salt) or 0 (pepper) with probability p.

Motion Blur from fast moving vehicles smears the intensities along the motion direction [3]. We simulate linear motion blur via convolution of the image with a box kernel of width W in either horizontal or vertical direction.

Defocus Blur from defocused cameras is simulated by convolving the image with a symmetric 2-D Gaussian kernel with standard deviation σ.

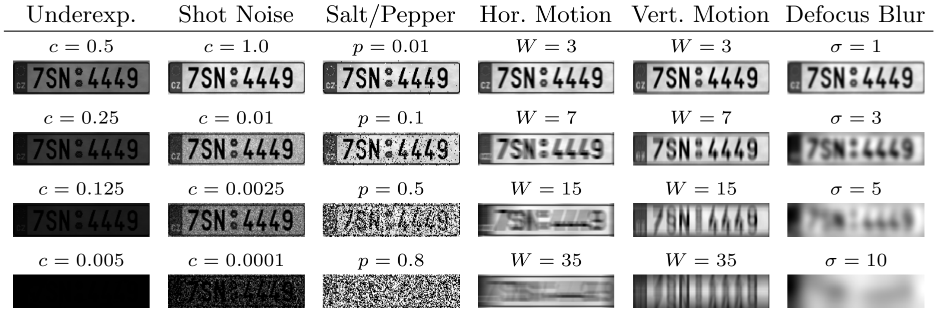


Fig. 2: Exemplary synthetic samples with increasing level of degradation per row.

4 Evaluation

We report accuracies for the correct detection of whole LP sequences. The top-1 accuracy of the proposed method with best-path decoding [5] (CRNN) is compared to Lorch *et al.* [11] (CNN). We also report CRNN top-n accuracies, which are not available for the CNN (cf. Sec. 5).

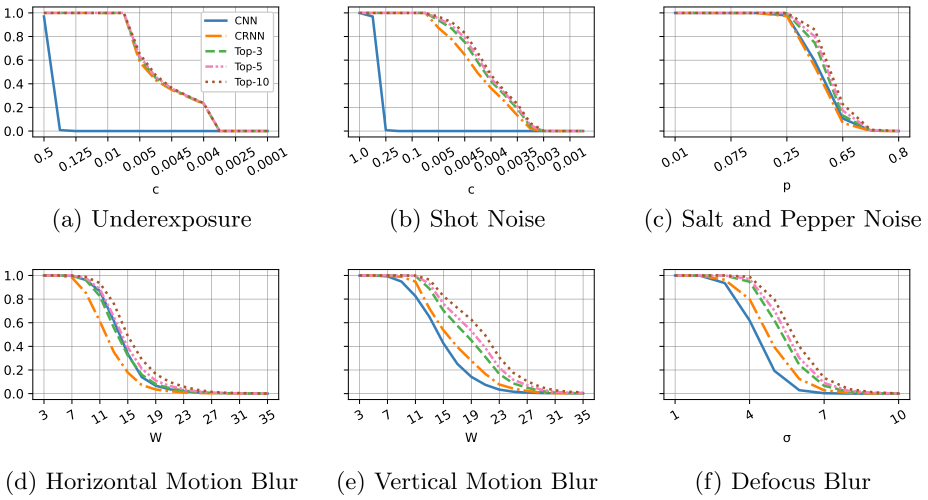
The CRNN top-n predictions are obtained from beam search decoding with beam sizes $n \in \{3, 5, 10\}$. This beam search performs a breadth-first search with best-first strategy that explores n label solutions per time step in a sequence.

4.1 Robustness Testing on Synthetic Out-of-Distribution Data

We evaluate the impact of each type of unseen degradation individually, except for shot noise, where the addition of Poisson noise comes with a reduction of exposure. Each individual unseen degradation parameter is evaluated using 10 000 test samples. Figure 2 shows example test samples with their associated degradation parameters.

Underexposure We chose 15 levels of underexposure with $c \in 10^{-3} \cdot \{$ 0.1, 1, 2.5, 3.75, 4, 4.25, 4.5, 4.75, 5, 7.5, 10, 100, 125, 250, 500$\}$. The values are chosen adaptively during evaluation to densely cover regions with major accuracy variations. Figure 3a shows that the CRNN with best path decoding clearly outperforms the CNN. The CRNN robustly handles underexposure unless $c < 0.0075$, while the CNN collapses for all $c \leq 0.25$. Beam search decoding does not improve performance over best path decoding.

Shot Noise Since the standard deviation of shot noise is determined in dependence of the exposure, we re-use the exposure parameter c and evaluate 18 adaptively chosen parameters $c \in 10^{-3} \cdot \{0.1, 1, 2.5, 3, 3.25, 3.5, 3.75, 4.4, 4.25, 4.5, 4.75, 5, 10, 100, 125, 250, 500, 1000\}$. Figure 3b shows the results. The CRNN model's performance using best path decoding outperforms the CNN. The CRNN accuracy is almost 1 for $c \geq 0.01$, and degrades steadily until $c = 0.003$. In contrast, the CNN performance drops sharply for $c \leq 0.5$. Compared to the previous experiment that only evaluates underexposure without noise, both



(a) Underexposure

(b) Shot Noise

(c) Salt and Pepper Noise

(d) Horizontal Motion Blur

(e) Vertical Motion Blur

(f) Defocus Blur

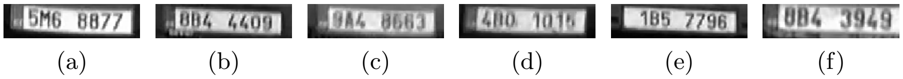
Fig. 3: Accuracy results of CNN and CRNN on synthetic test data sets. Detailed numerical results are available.[1]

models perform better. We assume that in this case, the additional Poisson noise slightly enhances the differences between the original pixel contrast. Beam search decoding leads to a further increase of the CRNN performance. This performance gain is already achieved with a beam width of $n = 3$, which indicates that the correct LP sequence is oftentimes only marginally missed in the top-1 prediction.

Salt and Pepper Noise Salt and pepper noise is added with increasing probabilities $p \in \{0.01, 0.05, 0.075, 0.1, 0.25, 0.5, 0.65, 0.75, 0.8\}$ until the model accuracy drops to 0. Figure 3c shows the results. The CNN and the CRNN with best path decoding are very robust until $p \geq 0.25$, and performance gently degrades for higher noise levels. For $p \geq 0.1$, the CNN marginally outperforms the CRNN, but beam search decoding provides a major performance gain. Already, the top-3 accuracy improves the CRNN performance over the CNN. The top-10 predictions further increase accuracy over the CNN, most notably for $p = 0.5$ (0.8584 versus 0.2663) and $p = 0.65$ (0.2276 versus 0.1212).

Horizontal Motion Blur We simulate 17 horizontal motion blur kernels with widths $W \in \{3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35\}$. Figure 3d shows the results. The CNN and the CRNN with best path decoding exhibit similar performance, with slightly better results for the CNN. Beam search decoding increases the performance of the CRNN already for $n = 3$. For beam width $n = 10$, the top-n accuracy is on average 0.0458 higher than the CNN. In practice, this increases the chance of finding the correct LP, even though it has to be identified from 10 candidates.

[1] https://www.cs1.tf.fau.de/research/multimedia-security/code/



(a) (b) (c) (d) (e) (f)

Fig. 4: Exemplary samples form the real world data set *ReId* [17].

Vertical Motion Blur To account for vertical motion blur, we use the same 17 motion blur kernels, but transpose each to flip it by ninety degrees. The results are shown in Fig. 3e. Overall, all accuracies are considerably higher, which is expected, as vertical motion does not smear across character boundaries, and also does not interfere with the CRNN's subdivision of the input into vertical slices. The CRNN achieves considerably higher accuracies than the CNN for all kernel sizes. For $W \geq 7$, the accuracy gently degrades. Beam search decoding boosts the CRNN performance. The average improvement over best path decoding is 0.0634, 0.0903 and 0.1260 for $n = 3$, $n = 5$ and $n = 10$.

Defocus Blur The test data set is degraded by defocus blur with $1 \leq \sigma \leq 10$. The results are shown in Fig. 3f. The CRNN with best path decoding surpasses the CNN's performance. For $4 \leq \sigma \leq 6$, the accuracy declines. Beam search decoding provides a notable performance increase. With $n = 10$, the accuracy remains above 0.1 for $\sigma \leq 7$.

4.2 Performance on Real World Data

We use the *ReId* dataset by Špaňhel *et al.* [17] to test on low-quality real-world images from Czech LPs. The dataset consists of 76 412 images from video cameras on highway bridges. Approximately 99.67% of the samples show Czech LPs, the remaining 0.03% originate from other European countries. Example images are shown in Fig. 4.

We add an eighth output layer to the CNN to accommodate for the 8 characters of Czech LPs as described by Kaiser *et al.* [7]. The CRNN architecture is used without any changes regarding the architecture. Without retraining, the CNN accuracy is 0.6806, and the CRNN accuracy is only 0.1346, which indicates larger difficulties to generalize to this dataset. Beam search achieves top-n accuracies of 0.2068, 0.2406 and 0.2849 for $n \in \{3, 5, 10\}$. The difficulties of both models to generalize is expected, even if most samples seem to be readable for human observers. The *ReId* dataset contains combined and unseen degradations like distortion and rotation. In addition, the LPs are not carefully aligned within the images.

Retraining the neural nets on the training part of the dataset of 105 924 images results in a significant improvement. The CNN achieves an accuracy of 0.9042, the CRNN achieves an accuracy of 0.9807 with best path decoding. The top-n accuracies are even slightly higher with 0.9847, 0.9867 and 0.9891 with beam sizes $n \in \{3, 5, 10\}$.

5 Discussion

The results show that the proposed CRNN with best path decoding outperforms the CNN on most out-of-distribution degradation situations. When testing on synthetic data, the CRNN is superior for underexposure, shot noise, vertical motion blur and defocus blur. The CNN performs slightly better on salt and pepper noise and horizontal motion blur. The advantage of the CRNN becomes more apparent when investigating the benefit of top-3 and top-5 predictions, which always outperform the CNN by a large margin. Without retraining, the CNN generalizes better to real-world data with unseen degradations. However, when fine-tuning the networks, the CRNN again outperforms the CNN.

The possibility to perform top-n predictions is particularly interesting for criminal investigations. For example, given 5 predictions, false candidates can be eliminated using additional knowledge like queries for registered LPs and the vehicle make and model. This benefit is only available for the CRNN: the CNN can only provide top-n predictions per character, not per sequence. Converting such top-n character predictions to reasonable top-n sequences is a challenge in its own right. Furthermore, the CRNN can predict arbitrary numbers of characters on a license plate, while the CNN must be adapted to each length.

6 Conclusion

Forensic investigations can benefit from the possibility to decipher LP images with severe degradations from unknown sources. In this work, we propose sequence learning on such strongly degraded LP images. The proposed CRNN outperforms the CNN approach by Lorch *et al.* [11] and Kaiser *et al.* [7] by a large margin. A particular benefit of the CRNN are the top-n most likely sequences. In many cases, the correct LP sequence is among the first few predictions. While the CRNN with only a single prediction already performs strongly, the top-n for $n \in \{3, 5, 10\}$ considerably improve the results. In forensic practice, it is very reasonable to work with such a relatively small number of ranked predictions.

Future work may further investigate deeper into horizontal motion blur, and consider perspective distortions and rotation. It may also be interesting to investigate the CRNN on LPs from countries like Germany that impose richer syntactic sequence rules.

References

1. Abdullah, S., Hasan, M.M., Islam, S.M.S.: YOLO-Based Three-Stage Network for Bangla License Plate Recognition in Dhaka Metropolitan City. In: IEEE International Conference on Bangla Speech and Language Processing. pp. 1–6 (2018)
2. Agarwal, S., Tran, D., Torresani, L., Farid, H.: Deciphering Severely Degraded License Plates. Electronic Imaging **2017**(7), 138–143 (2017)
3. Bovik, A.C.: Handbook of Image and Video Processing. Academic press (2010)

4. Council of EU: Collection of Laws of the Czech Republic. No. 343/2014 (2014)
5. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In: 23rd International Conference on Machine Learning. pp. 369–376 (2006)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. In: IEEE International Joint Conference on Neural Networks. vol. 4, pp. 2047–2052 (2005)
7. Kaiser, P., Schirrmacher, F., Lorch, B., Riess, C.: Learning to Decipher License Plates in Severely Degraded Images. In: Pattern Recognition. ICPR International Workshops and Challenges. pp. 544–559. Springer International Publishing (2021)
8. Laroca, R., Menotti, D.: Automatic License Plate Recognition: An Efficient and Layout-Independent System Based on the YOLO Detector. In: Anais Estendidos do XXXIII Conference on Graphics, Patterns and Images. pp. 15–21. SBC (2020)
9. Laroca, R., Severo, E., Zanlorensi, L.A., Oliveira, L.S., Gonçalves, G.R., Schwartz, W.R., Menotti, D.: A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector. In: IEEE International Joint Conference on Neural Networks. pp. 1–10 (2018)
10. Li, H., Wang, P., Shen, C.: Towards End-to-End Car License Plates Detection and Recognition with Deep Neural Networks. IEEE Transactions on Intelligent Transportation Systems **20**(3), 1126–1136 (2018)
11. Lorch, B., Agarwal, S., Farid, H.: Forensic Reconstruction of Severely Degraded License Plates. Electronic Imaging **2019**(5) (2019)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once:Unified, Real-Time Object Detection. In: IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
13. Riaz, W., Azeem, A., Chenqiang, G., Yuxi, Z., Khalid, W., et al.: YOLO Based Recognition Method for Automatic License Plate Recognition. In: IEEE International Conference on Advances in Electrical Engineering and Computer Applications. pp. 87–90 (2020)
14. Rossi, G., Fontani, M., Milani, S.: Neural Network for Denoising and Reading Degraded License Plates. In: Pattern Recognition. ICPR International Workshops and Challenges. pp. 484–499 (2021)
15. 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 Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2298–2304 (2016)
16. Shivakumara, P., Tang, D., Asadzadehkaljahi, M., Lu, T., Pal, U., Anisi, M.H.: CNN-RNN based method for license plate recognition. CAAI Transactions on Intelligence Technology **3**(3), 169–175 (2018)
17. Špaňhel, J., Sochor, J., Juránek, R., Herout, A., Maršík, L., Zemčík, P.: Holistic Recognition of Low Quality License Plates by CNN using Track Annotated Data. In: 14th IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 1–6 (2017)
18. Suvarnam, B., Ch, V.S.: Combination of CNN-GRU Model to Recognize Characters of a License Plate number without Segmentation. In: 5th International Conference on Advanced Computing & Communication Systems. pp. 317–322 (2019)
19. Tourani, A., Shahbahrami, A., Soroori, S., Khazaee, S., Suen, C.Y.: A Robust Deep Learning Approach for Automatic Iranian Vehicle License Plate Detection and Recognition for Surveillance Systems. IEEE Access **8**, 201317–201330 (2020)
20. Zhang, H., Sun, F., Zhang, X., Zheng, L.: License Plate Recognition Model Based on CNN + LSTM + CTC. In: International Conference of Pioneering Computer Scientists, Engineers and Educators. pp. 657–678. Springer (2019)