Neurocomputing 449 (2021) 189–206

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

V-LPDR: Towards a unified framework for license plate detection, tracking, and recognition in real-world traffic videos

Cong Zhang, Qi Wang *, Xuelong Li

School of Computer Science and Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, 710072 Shaanxi, PR China

ARTICLE INFO

Article history:
Received 2 June 2020
Revised 16 January 2021
Accepted 26 March 2021
Available online 01 April 2021
Communicated by Zidong Wang

Keywords:
License plate detection
License plate recognition
Deep learning

ABSTRACT

License plate detection and recognition (LPDR) has attracted considerable attention in recent years, and many algorithms have presented the competitive performance on several datasets. However, there are still three significant issues to be addressed in this field. Firstly, most methods have poor detection performance in unconstrained scenarios with moving vehicles and highly distracting background objects. Secondly, existing systems generally focus on single image-based algorithms, yet traffic video sequences provide more effective information than individual frames for LPDR tasks. Thirdly, images and videos captured in complex environments may be adversely affected by distortions and low resolution, causing sensitive recognition performance and reduced robustness. To remedy these issues, we propose to automatically perform license plate detection, tracking, and recognition in real-world traffic videos and integrate them into a unified end-to-end framework via deep learning. The contributions of this paper are threefold: 1) A deep flow-guided spatiotemporal license plate detector is proposed to model the video contextual information by introducing optical flow and a novel spatiotemporal attention mechanism; 2) An online license plate tracker is developed to bridge video-based detection and recognition which utilizes both motion and deep appearance information, and innovatively, it can be end-to-end trained with the detector via multi-task learning; 3) The efficient quality-guided license plate recommender and recognizer are proposed to jointly perform stream recognition. The former recommends high-quality frames from video streams while the latter generates recognition results. We evaluate the proposed method on three traffic video-based license plate datasets, and ablation studies have been presented to verify the effectiveness of each component mentioned above. Moreover, extensive experiments are conducted for comparison with other approaches in different scenarios, and the results have demonstrated that our method achieves state-of-the-art performance on all datasets.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Recently, Intelligent Transportation Systems (ITS) have drawn critical attention due to their vital roles in smart cities and autonomous driving [1]. ITS aims to efficiently manage the privately-owned vehicles, the number of which has increased significantly in the past years. It also brings tremendous convenience and security to the users along with the development of related intelligent industrial technology [2,3]. As an indispensable component of ITS, automatic license plate detection and recognition (LPDR) can greatly alleviate traffic management burdens, which has attracted considerable research interests [4]. Furthermore, the technology LPDR can also be applied in many real-world scenarios, such as parking lot access control, toll collection, and traffic law enforcement [5].

In general, a license plate is composed of six or seven characters with various fonts, colors, backgrounds, and aspect ratios in different regions. Over the past decades, numerous LPDR algorithms based on computer vision have been extensively explored and studied to deal with these challenges. Moreover, diverse datasets have been released publicly, such as Caltech-cars (Real) 1999 dataset [6], PKU vehicle dataset [7], Application-Oriented License Plate (AOLP) dataset [8], Chinese City Parking Dataset (CCPD) [9], SSIG-SegPlate dataset [10], and UFPR-ALPR dataset [11]. Some proposed approaches have achieved state-of-the-art performance on several public LPDR datasets. For instance, the method introduced in [12] presented an average recognition accuracy of 95.83% on AOLP dataset and an overall recognition accuracy of 98.9% on CCPD dataset.

* Corresponding author.
E-mail addresses: nwpuzhangcong@gmail.com (C. Zhang), crabwq@gmail.com (Q. Wang), xuelong_li@nwpu.edu.cn (X. Li).

https://doi.org/10.1016/j.neucom.2021.03.103
0925-2312/© 2021 Elsevier B.V. All rights reserved.

Most previous studies usually perform on the low-level handcrafted features [5,13], causing poor performance in highly complicated and unconstrained environments. Thanks to recent development and advances of deep learning in many fields like image processing [14–17], convolutional neural networks (CNNs) bring the community inspiration to accomplish the LPDR task in complex and changeable scenarios. Furthermore, based on the high-level discriminative features extracted by CNNs, recurrent neural networks (RNNs), especially long short-term memory (LSTM) can be implemented to read the characters and numbers in the license plates. Both CNNs and RNNs demonstrate the impressive performance for enhancing the comprehension of images, which have been adopted in LPDR systems. For the first time, Li et al. [18] treated the plate as a string of text and tackled LPDR by leveraging the high capability of CNNs and LSTM to render it suitable in open environments. After that, more advanced deep learning-based methods have been proposed consecutively, which also obtain competitive results on different datasets [12,19–23].

Both traditional and deep learning-based LPDR algorithms have been widely studied for many years. Most existing methods involve expensive and strict equipment such as sophisticated image capture devices, fixed rotation angles, or controlled translation. However, it is difficult to meet the above requirements in real-world scenarios. As illustrated in Fig. 1, different interference like similar background objects, occlusion, motion blur, and non-uniform illumination will unavoidably occur under such dynamic and unconstrained conditions. Consequently, these algorithms may not work well in the complex environments.

Although some solutions have been specially introduced to perform the LPDR task in complex and open scenarios [12,19,24–27], there are still limitations and deficiencies in real-world traffic applications. To sum up, most existing LPDR algorithms in the literature have three main problems: (1) Poor detection performance on single images with highly complicated backgrounds and motion blur, leading to reduced availability in complex scenes such as intelligent driving. (2) Only perform license plate detection and recognition on the individual images, while it is more worthy of attention to involving detection, tracking, and recognition in traffic videos instead of the single images. (3) Sensitive recognition performance on the real-world images or videos captured in harsh environments and suffer from low resolution and illumination variation.

Against the above issues, this work sheds new light on the unified video-based LPDR systems. To be specific, in this paper, we propose a novel deep learning-based framework that is capable of automatically detecting, tracking and recognizing license plates in the dynamic, unconstrained and complex scenarios. To the best of our knowledge, it is the first work to unify license plate detection, tracking, and recognition into one computational framework via deep learning. Moreover, considering the availability and popularity of video clips, we tackle the LPDR task in real-world traffic videos rather than the single image-based methods. It is worth noting that several existing approaches do claim to present the video-based LPDR systems [11,28–30], yet they merely focus on majority voting or super-resolution during the recognition stage. Nevertheless, in this work, not only a compact and efficient multi-frame based recognition network is developed, but also the spatiotemporal detection and online tracking are explored by making the best of temporal contextual knowledge. The whole unified **V**ideo-based **LPDR** framework, namely V-LPDR, which can perform license plate detection, tracking, and recognition, is the main contribution of this paper. As depicted in Fig. 2, V-LPDR consists of three components, corresponding to its three subtasks, license plate detection (LPD), license plate tracking (LPT), and license plate recognition (LPR). For clear demonstration, the above three components are summarized and introduced as follows.

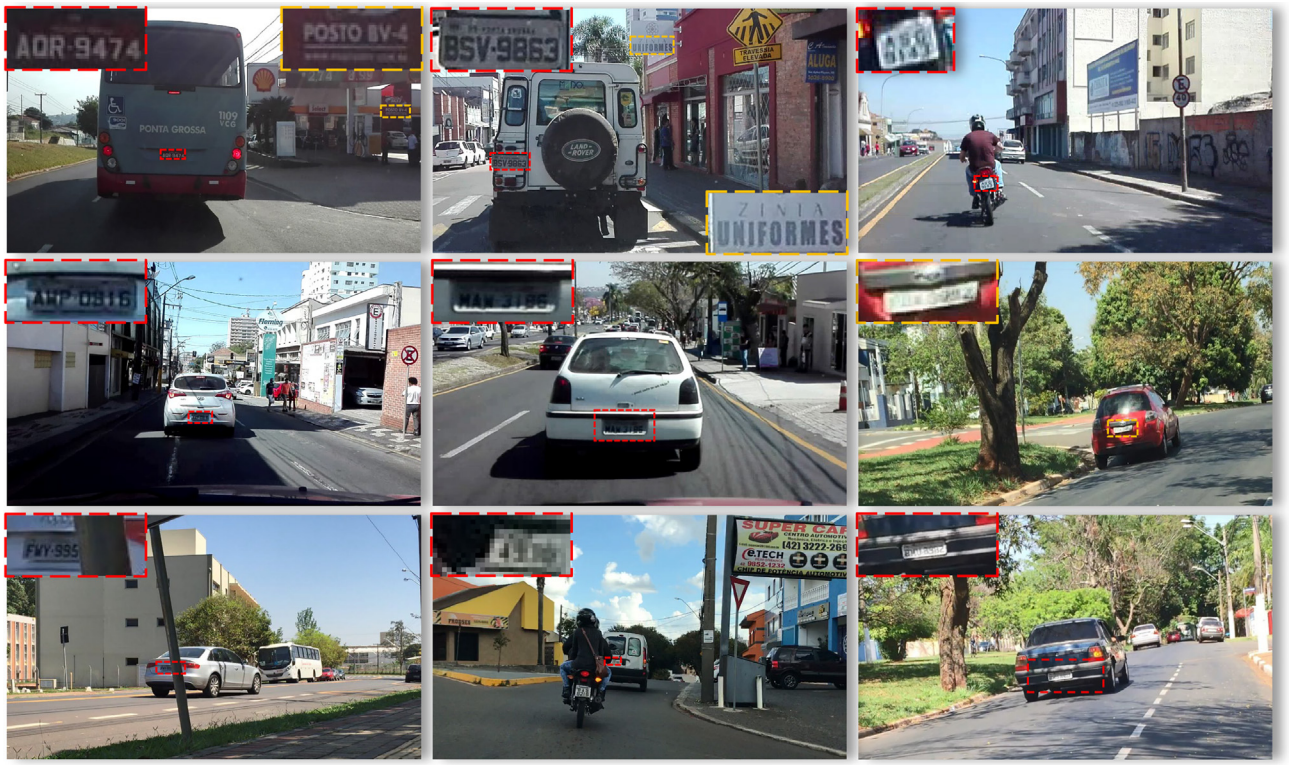

Fig. 1. Illustration of diverse inference for LPDR systems in real-world scenarios, such as background objects, perspective distortion, motion blur, occlusion, and illumination variance.

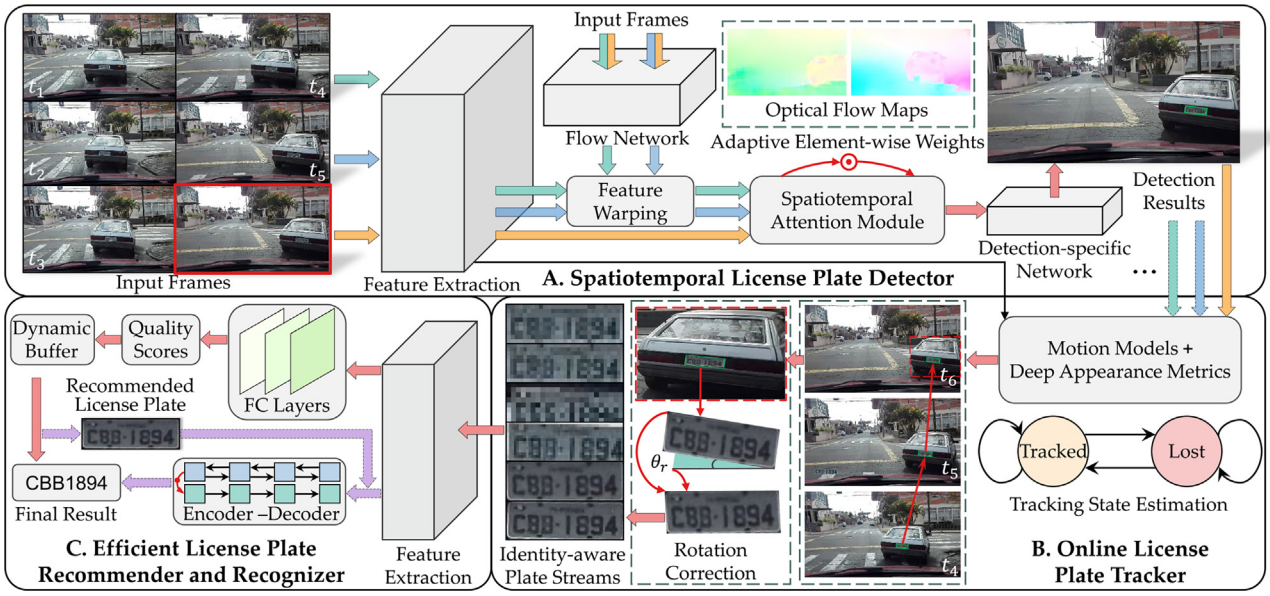

Fig. 2. Overview of the proposed video-based LPDR framework. The blue, green and yellow arrows in the detector represent the processing flow of three different frames t_4, t_5 and t_6. The frame t_6 is regarded as the current frame in the illustration.

1.1. Spatiotemporal license plate detector

A novel deep spatiotemporal detection network is developed for the subtask LPD that is the first and important stage of V-LPDR. As shown in Fig. 2, having aggregated the optical flow maps of adjacent frames, the proposed license plate detector utilizes rich temporal contextual information in the videos to improve the detection performance and robustness despite motion blur and perspective distortion. Furthermore, by introducing the attention module to learn the spatial distribution of license plates, the detector can achieve high detection precision in the complex scenarios.

1.2. Online license plate tracker

As to the second subtask LPT, an online tracking algorithm is adopted based on the deep local appearance features. Unlike previous LPDR approaches that usually ignore the tracking stage, V-LPDR employs both motion models and deep appearance metrics to track the detected plates. Moreover, in the proposed framework, license plate tracking is innovatively integrated with detection to reduce the additional computational complexity of tracking features generation. The tracker plays a pivotal role in V-LPDR due to its selection for detection and guidance on recognition.

1.3. Efficient license plate recommender and recognizer

It is not robust and accurate to derive the recognition results from the individual frames with distortions, which significantly limits the real-world applications of LPDR. Considering both computational efficiency and the improvement of multi-frame recognition, a novel lightweight recognition network is proposed, as illustrated in Fig. 2, which is composed of the recommender and recognizer. With the high-quality frames recommended by the former, the latter can obtain satisfactory results in unconstrained and complex scenarios.

The rest of this paper is organized in the following way. A brief review of related work is given in Section 2. Section 3 describes the proposed framework in detail. In Section 4, the experimental results on different datasets are shown and analyzed to verify the effectiveness of our approach. Finally, we summarize this paper and present the future work on LPDR systems in Section 5.

2. Related work

With great application potential, the LPDR system plays an essential role in ITS. Despite the fact that a considerable amount of academic approaches have been published over the past two decades, most of them only concentrate on each individual image while few video-based LPDR methods explicitly employ temporal information of video sequences [31]. In general, the typical single image-based LPDR task is addressed to two key components, *i.e.* license plate detection and recognition. As to video-based LPDR, several techniques are utilized to obtain accurate recognition results. For a clear introduction, in this section, the literature overview is provided in three aspects: license plate detection, license plate recognition, and video-based techniques for LPDR systems.

2.1. License plate detection

License plate detection aims to discover and locate all plates in the single image or video frame, and then the bounding boxes enclosing the predicted potential regions are generated in this stage. In terms of the significant improvement of LPD by deep learning in recent years, existing algorithms can be roughly divided into two categories [5,26,27]: *traditional handcrafted features*-based and *deep learning*-based methods.

Traditional handcrafted feature-based methods generally localize the plates using boundary information, color features, and texture features [5]. As license plates in almost all countries have the rectangle shapes with their known aspect ratios, boundary or edge information can be exploited to find potential rectangles in the whole image. For example, in [32], Sobel filter was employed to detect boundaries based on the color transition between the plate and bodywork. Recently, Yuan et al. [33] proposed a novel line density filter to extract candidate regions in a binary edge image. Boundary information-based approaches are fast and simple, yet they are unsuitable for complex images and videos because of the sensibility to unwanted boundaries. In line with the observation that license plates usually have the specific color combination

of backgrounds and characters, color features-based methods are typically traditional algorithms for LPD. A new method depending on template matching by analyzing target color pixels was proposed in [34]. With periodic strip search, this method finds the hue of each pixel, while it is designed for Iranian plates only, leading to great limitations on other regions. Texture feature-based approaches utilize the unconventional pixel intensity distribution in the plate regions. In [35], Yu et al. adopted empirical mode decomposition (EMD) analysis to find the desired wave crest that pointed out the position of a plate after obtaining the vertical and horizontal details. In [36], a scale-adaptive deformable part-based model was proposed which selected the most prominent features and reduced the detection time by avoiding the evaluation at different scales. Both color feature- and texture feature-based approaches can detect plates with deformed or damaged edges. However, they may not work well in the case of complicated backgrounds and various non-uniform illumination.

Deep learning-based LPD methods have received considerable attention after the advancement of deep learning in the field of object detection. Over the last couple of years, some impressive algorithms are proposed to address the problem of plate detection using deep CNNs [18,19,21,22]. For the first time, Li et al. [18] exploited a well-trained 31-class CNN to detect characters, followed by a plate/non-plate CNN classifier to eliminate false positives. Unfortunately, with many pre- and post-processing modules, such a complex model cannot be trained in an end-to-end manner, which limits its speed and performance. In [23], Chen et al. proposed a fully convolutional network for random-positioned detection by the fusion of multi-scale and hierarchical features, but it was only designed for Macau license plates. Considering the computational efficiency in real-world applications, the real-time LPD algorithms [4,11,37–39] were proposed based on YOLO detector. In [4], Xie et al. developed a YOLO-based model to predict the coordinates and rotation angles of plates in addition to their confidence values. Whereas, their approach can only be forced to output one bounding box in each frame, which makes it inappropriate in complex scenarios with multiple vehicles. Henry et al. [39] proposed the modified tiny YOLOv3 for plate localization and then utilized license plate layout detection to improve the multinational character recognition accuracy. In order to reduce the interference of complicated backgrounds, several recent works [11,24,28,37] perform vehicle detection before plate detection, but causing additional operational overhead.

2.2. License plate recognition

License plate recognition is intended for reading the numbers and characters from the detected candidate regions. In general, the technology used in this stage is optical character recognition (OCR). With the development of sequence-to-sequence model in the field of scene text spotting [40,41], most existing LPR algorithms can be categorized into two types: *segmentation-based* and *segmentation-free* methods.

Segmentation-based methods consist of two separate phases, *i.e.*, license plate segmentation and character recognition, which are commonly applied in previous LPDR systems. For example, Hsu et al. [42] first adopted the maximally stable extremal region (MSER) to segment plates and then the local binary pattern (LBP) features were extracted and classified by introducing a linear discriminant analysis (LDA) classifier for character recognition. Since plate segmentation by itself is a really challenging task [19], there are many works dedicated to it [20,43]. For example, in [43], the optimal K-means (OKM) clustering-based segmentation method was introduced to segment license plate characters, which could deal with the rotated plates. As to character recognition, template matching-based and machine learning-based methods are usually used for the segmented individual characters. Template matching-based methods [44] recognize each character in terms of the similarity measure between templates and characters. Artificial neural networks (ANN) [45,46] and support vector machine (SVM) [47] are adopted as machine learning-based methods for recognition after feature extraction. In addition, inspired by the impressive performance of extreme learning machines (ELM) [48–50] on classification, some researchers [51,52] exploited ELM as a classifier for the task of license plate character recognition, which yielded competitive results. By introducing segmentation, plate recognition can be simplified as the classification tasks, yet most segmentation-based methods fail to segment plates with adhesion and distorted characters.

Segmentation-free methods avoid the error accumulation for recognition caused by segmentation failures, which have received considerable research attention of the LPDR community. In [53], segmentation and character recognition were jointly performed utilizing hidden Markov models (HMMs). Inspired by scene text recognition, Li et al. [18] viewed plates as strings of text and exploited RNNs followed by connectionist temporal classification (CTC) for plate sequence recognition without the character-level segmentation. Nonetheless, Dong et al. [54] claimed that the proposed approach in [18] was fragile to distortions in the wild. In [12], Zhang et al. employed the CycleGAN model for plate generation and a 2D attentional-based image-to-sequence network for plate recognition, which avoided character segmentation. However, faced with motion blur and distortions, most of these algorithms cannot achieve satisfactory recognition performance in the individual images. Against this issue, we attempt to perform segmentation-free recognition in the multiple consecutive frames instead of the single frame.

2.3. Video-based techniques for LPDR systems

Despite an enormous amount of academic literature on the LPDR topic, only a few video-based algorithms explicitly adopt temporal contextual information [31]. Note that most existing methods still work on single images even though it is mentioned to extract each frame from the video sequences. There are three typical techniques or strategies for video-based LPDR systems: *license plate tracking*, *majority voting*, and *super-resolution*.

License plate tracking aims to link the positioned plates in consecutive frames and then generate the trajectory. It is worth noting that plate tracking is rarely studied separately in the literature [55]. On the contrary, it is commonly implemented for other tasks such as plate detection, plate recognition and vehicle speed measurement. Voted block matching [56], differential evolution [57], Kalman filter [58], MSER [59], and Kanade-Lucas-Tomas (KLT) algorithm [60,61] have been developed and exploited to perform plate tracking in the time sequences of videos. Recently, Luvizon et al. [61] proposed a nonintrusive video-based system, which first localized plates in image regions and then tracked them using the pyramidal KLT algorithm across multiple frames, and vehicle speed could be finally measured by comparing the trajectories.

Majority voting was utilized to merge the high-level recognition outputs and generate the final decision for each plate character [11,28]. Depending on the performance of segmentation, these methods may not be robust especially in complex dynamic scenarios. In addition to majority voting, *super-resolution* is intended to generate a single high-resolution image from multiple low-resolution observation sequences. License plate super-resolution can enhance the recognition accuracy, which has attracted great research interests in recent years [62,63,29,30,64]. For example, in [62], Zhang et al. introduced a multi-task generative adversarial network (GAN) for joint plate super-resolution and recognition. However, super-resolution requires extra computing resources,

leading to a compromise of its effectiveness. Moreover, the latter two video-based techniques only occupy the subtask LPR without considering the impact of detection and tracking on the recognition inputs.

3. Our approach

Different from the methods mentioned above, we perform the task of license plate detection, tracking, and recognition in complex and unconstrained traffic videos. The whole proposed framework is illustrated in Fig. 2, which consists of three components, spatiotemporal license plate detector, online license plate tracker, and efficient license plate recommender and recognizer, corresponding to the three subtasks. Concretely, with an input video, the **detector** is first responsible for localizing plates by referring to the temporal relationship among multiple adjacent frames and spatial information in the current frame. Next, the **tracker** can generate plate streams and assign them different identities using motion information and discriminative features. Finally, the **recommender** is developed to score each stream, and the plate region with the highest quality score will be selected for the identity-aware **recognizer**, where the recognition prediction represents the final results of its corresponding stream. Note that the proposed V-LPDR is a completely online framework, which means that it depends on the past information without any future frames in the videos.

In this section, we will provide detailed descriptions of the above three components respectively.

3.1. Flow-guided spatiotemporal attention detection network

In order to detect license plates in complex and unconstrained scenarios, we present a novel flow-guided spatiotemporal attention network for the subtask LPD. The proposed detection architecture is depicted in Fig. 3. It aims at dealing with complex situations such as drastic appearance changes, motion blur, and part occlusion, as exemplified in Fig. 1. On the one hand, inspired by [65–67], the feature aggregation of neighboring frames is exploited to achieve high detection recall based on temporal attention and optical flow information. On the other hand, observing complicated background interference in real-world traffic videos, spatial attention mechanism is also adopted for high-level discriminative features to improve the precision. As shown in Fig. 3, the proposed network can be divided into three modules: detection backbone, flow-guided feature warping, and spatiotemporal attention block. An efficient scene text detector in [40] is adapted as our plate detection backbone which consists of the feature extraction subnetwork N_e and the detection-specific subnetwork N_d.

3.1.1. Flow-guided temporal feature warping

Formally, given a series of input video frames $\{I_t\}$, where $t \in \{1, \cdots, T\}$ and T represents the total number of frames in the input video. The proposed detector is intended for generating all potential plate bounding boxes R_t for each frame I_t. There is rich semantic information in high-level deep features which can enhance detection performance. In view of that, with the r-th frame I_r in $\{I_t\}$, the well-known ResNet [68] is modified as N_e in our backbone to produce the feature maps $F_r = N_e(I_r)$. In the following description, we define I_r as the reference frame for convenience. However, such a rough feature map F_r involves no temporal coherence, causing its vulnerability and weak response to motion blur and video defocus. Against this issue, the motion-guided warping strategy is exploited to introduce temporal information in the adjacent frames of I_r, exemplified as I_i and I_j, where $i, j < r$ for online applications of LPDR. Let F_i and F_j be the deep features of I_i and I_j respectively, *i.e.* $F_i = N_e(I_i), F_j = N_e(I_j)$. In addition, as depicted in Fig. 3, their corresponding optical flow maps can be estimated by the flow network N_f [69], given by:

$$f_{r,i} = N_f(I_r, I_i),\ f_{r,j} = N_f(I_r, I_j). \tag{1}$$

The flow-guided warping mechanism is further applied to a pair of feature maps and flow maps, formulated as:

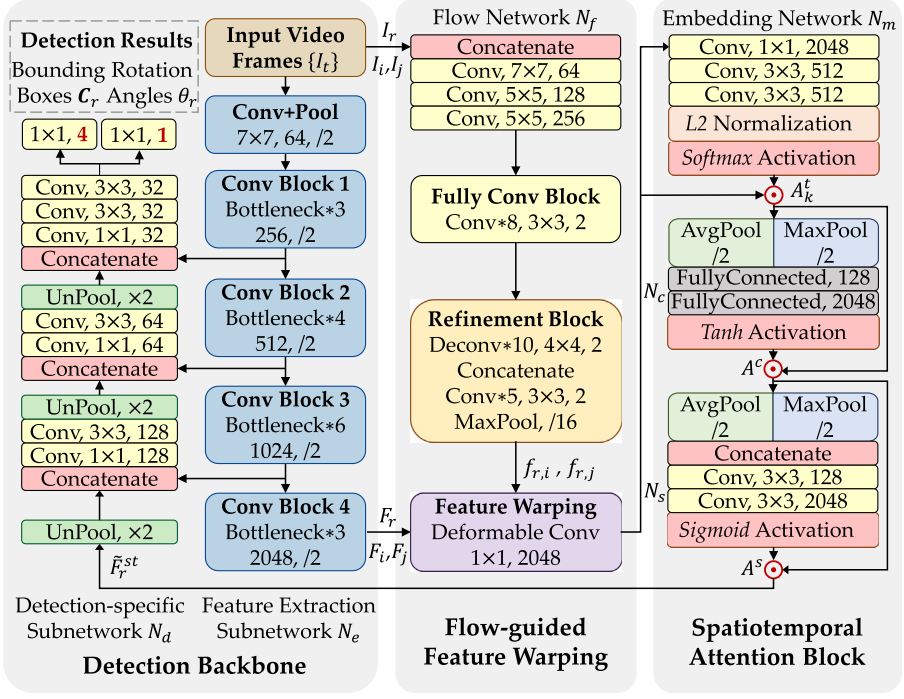

Fig. 3. The structure of the proposed license plate detection network. For convolution operations, the two parameters mean kernel size and the number of output channels, respectively. For pooling and unpooling operations, /2 and ×2 indicate halving or doubling the size of feature maps.

$$F_i^{warp} = Warp(F_i, f_{r,i}),$$
$$F_j^{warp} = Warp(F_j, f_{rj}). \tag{2}$$

where $Warp(\cdot)$ is the element-wise bilinear warping function for all locations in F_i and F_j, while F_i^{warp} and F_j^{warp} denote their warped feature maps respectively.

3.1.2. Spatiotemporal attention

After the flow-guided feature warping, F_r, F_i^{warp}, and F_j^{warp} will be aggregated and refined into one feature map based on the novel spatiotemporal attention. It is composed of three steps: temporal feature aggregation, channel-wise feature clarification, and spatial location refinement. For temporal aggregation, given the superset $\mathscr{F} = \left\{F_r, F_i^{warp}, F_j^{warp}\right\}$, different adaptive attention weights are generated for each element. Following [65], the measurement features are first generated for similarity comparison by applying the tiny 3-layer embedding subnetwork shown in Fig. 3, formulated as $\left\{F_r^{ms}, F_i^{ms}, F_j^{ms}\right\} = N_m(\mathscr{F})$. Then the temporal attention weights can be estimated by:

$$A_k^t = \frac{exp\left(F_k^{ms} \odot F_r^{ms}\right)}{\sum_{k\prime\in\{i,j\}} exp\left(F_{k\prime}^{ms} \odot F_r^{ms}\right)}, \tag{3}$$

where $k \in \{i,j\}$ and $\odot$ means the element-wise multiplication. Specifically, softmax function is utilized for normalizing each element to the range of $[0,1]$. Given the channel c of $A_k^t, A_k^t(c)$ is a 2-D weight matrix instead of a scalar, constrained by $\sum_{k\in\{i,j\}} A_k^t(c) = \mathbb{1}\mathbb{1}^{\mathrm{T}}$. The temporal aggregated feature for the reference frame I_r can be further computed by:

$$\widetilde{F}_r^{tmp} = F_r + \sum_{k\in\{i,j\}} A_k^t \odot F_k^{warp}. \tag{4}$$

In practice, A_k^t is channel-independent while $\widetilde{F}_r^{tmp}$ involves the contextual information across consecutive frames. Such temporal attention focuses on the motion model in the flow fields like moving vehicles but without learning the spatial distribution of license plates in the input frame.

In order to localize the plate regions with high precision and robustness rather than diverting attention to the entire moving vehicles, spatial attention is also developed after temporal aggregation. Specifically, motivated by [70–72], we introduce the channel-wise feature attention followed by spatial location attention. Formally, given the temporal aggregated feature maps $\widetilde{F}_r^{tmp} \in \mathbb{R}^{H\times W\times C}$, the channel-wise feature attention map $A^c \in \mathbb{R}^{1\times 1\times C}$ and spatial location attention map $A^s \in \mathbb{R}^{H\times W\times 1}$ are generated sequentially. As illustrated in Fig. 3, the channel-wise attention module consists of two different pooling operations $AvgP(\cdot)$ and $MaxP(\cdot)$ followed by the 2-layer fully connected network N_c. This module is designed for feature refinement by weighing the importance of channels in $\widetilde{F}_r^{tmp}$, formulated as:

$$A^c = tanh\left(N_c\left(AvgP\left(\widetilde{F}_r^{tmp}\right)\right) + N_c\left(MaxP\left(\widetilde{F}_r^{tmp}\right)\right)\right), \tag{5}$$

$$\widetilde{F}_r^{ch} = A^c \odot \widetilde{F}_r^{tmp}, \tag{6}$$

where $tanh(\cdot)$ means the non-linear activation function.

With the refined features $\widetilde{F}_r^{ch}$ by temporal aggregation and channel-wise attention, the fully convolutional network is developed to perform spatial location attention and explore the inter-spatial feature relationships. In general, the spatial location attention module aims to learn where the license plates are and where to attend in the whole individual images, which is complementary to the temporal attention. In line with the above hypothesis, denote N_s as the spatial attention generation network in Fig. 3 and its attention map can be generated by:

$$A^s = sigmoid\left(N_s\left(\left[AvgP\left(\widetilde{F}_r^{ch}\right); MaxP\left(\widetilde{F}_r^{ch}\right)\right]\right)\right), \tag{7}$$

$$\widetilde{F}_r^{st} = A^s \odot \widetilde{F}_r^{ch}, \tag{8}$$

where $sigmoid(\cdot)$ represents the sigmoid activation function while $[\cdot;\cdot]$ means the concatenation operation on the channel axis. It should be argued that these attention modules are lightweight with negligible overheads and end-to-end trainable. Moreover, as depicted in Fig. 3, $\widetilde{F}_r^{st}$ will be fed into the detection-specific subnetwork N_d to produce the final results, formulated as $\widetilde{F}_r = N_d\left(\widetilde{F}_r^{st}\right)$. The rotated bounding boxes $\mathscr{R}_r = \{\mathscr{C}_r, \theta_r\}$ enclosing the candidate plate regions can be further regressed through the detection backbone, described by the box coordinates $\mathscr{C}_r$ and their rotation angles θ_r.

3.2. Deep appearance representation-based online tracking

For video-based LPDR systems, plate tracking forms a connecting link between detection and recognition. Exactly speaking, it is conducted to group corresponding license plates into plate streams, as shown in Fig. 2. Intuitively, an available plate tracker ensures a similar metric distance inside each stream. Therefore, there are two important requirements: (1) The tracking metric should be robust with a high tolerance for diverse interference in complex real-world traffic videos. It would be better to learn license plate discrimination from deep feature representations. (2) The tracker's computational complexity should be decent without loss of effectiveness.

Inspired by that the learning-based appearance feature plays a constructive role in data association of multiple object tracking (MOT) [73,74], we solve the assignment problem for plate tracking by integrating the motion model with deep discriminative features. It should be noted that we mainly focus on the idea that deep appearance information is beneficial to tracking performance and there are also two crucial differences between the proposed plate tracker and Deep SORT algorithm in [74]. Firstly, in the online inference or tracking stage, the appearance feature vectors are generated in different ways. Our online tracker can produce the local feature vectors of plate regions based on the global feature map obtained from the proposed detection network, while Deep SORT directly adopts appearance features well-trained offline on another dataset instead of generating them online. Secondly, during the training stage, different training strategies are utilized to discriminate deep appearance features for tracking. Unlike that an extra appearance descriptor should be trained separately offline in Deep SORT, we innovatively propose to jointly training plate detection and tracking in this paper. Specifically, to reduce the computing resource for the generation of discriminative appearance features, its training process for tracking is combined into a multi-task architecture with our proposed detection network. As illustrated in Fig. 4, these two tasks can be jointly trained via deep cosine metric learning and multi-task learning.

3.2.1. Joint metric for online plate tracking

In general, existing methods tackle the tracking task by associating the detection results across several neighboring frames, which can be regarded as tracking-by-detection paradigms. We develop the online plate tracking algorithm following this way. Having obtained the plate detections in Section 3.1, the recursive Kalman filter and data association are employed based on the sin-

gle hypothesis tracking approaches. In this paper, the standard Kalman filter is implemented with the linear observation model for state estimation, while data association aims at determining whether to steady the predicted plate state as *tracked* or update it to *lost*. For data association in complex traffic scenarios with uncalibrated cameras, we introduce the joint metric based on both motion information and deep appearance representations.

Algorithm 1. Online Matching for Plate Stream Generation

Input: Plate tracklets
$\mathscr{T} = \left\{\mathscr{T}^{(1)}, \cdots, \mathscr{T}^{(I_{max})}\right\} = \left\{\boldsymbol{t}_m^{(i)} | i \in \{1, \cdots, I_{max}\}, m \in \{1, \cdots, M\}\right\}$
and plate detections $\mathscr{D} = \{\boldsymbol{d}_n | n \in \{1, \cdots, N\}\}$
Initialize the tracking identities $ids \leftarrow \{1, \cdots, I_{max}\}$
Initialize the identity-aware license plate streams
$\widetilde{\mathscr{T}} \leftarrow \varnothing$
Initialize the unmatched plate detections $\widetilde{\mathscr{D}} \leftarrow \mathscr{D}$
for $n \in \{1, \cdots, N\}$ **do**
  for $i \in \{1, \cdots, I_{max}\}$ **do**
    for $m \in \{1, \cdots, M\}$ **do**
      Compute the joint metric $\widetilde{\mathscr{M}}_{n,m}$ using Eq. 11
    end for
    Save the minimal cost matching
  $\left\{\widetilde{\mathscr{M}}_{n,m\prime}, \boldsymbol{t}_{m\prime}^{(i)}, \boldsymbol{d}_n\right\}$
end for
  Compute the tracking state $\mathscr{S}_n$ using Eq. 12
  if $\mathscr{S}_n = tracked$ **then**
    Search and produce the optimal matching $\mathscr{T}^{(i^*)}, \boldsymbol{d}_n\}$
    Online Update tracking identities $ids \leftarrow ids \cup i^*$
    Online Update $\mathscr{T}^{(i^*)} \leftarrow \mathscr{T}^{(i^*)} \cup \boldsymbol{d}_n$,
  $\mathscr{T} \leftarrow \left\{\mathscr{T}^{(i)} | i \in ids\right\}$
  end if
  Online Update $\widetilde{\mathscr{D}} \leftarrow \widetilde{\mathscr{D}} \setminus \boldsymbol{d}_n$
end for
Update $\widetilde{\mathscr{T}} \leftarrow \left\{\left\{id, \mathscr{T}^{(id)}\right\} | id \in ids\right\}$
Output: The generated identity-aware plate streams $\widetilde{\mathscr{T}}$

Concretely in the frame I_r, denote the n-th axis-aligned bounding box detection by $\boldsymbol{d}_n$ and the m-th track by $\boldsymbol{t}_m$. Then the squared Euclidean distance is exploited to incorporate motion information, formulated as:

$$\mathscr{M}_{n,m}^{(1)} = (\boldsymbol{d}_n - \boldsymbol{t}_m)^{\mathrm{T}} (\boldsymbol{d}_n - \boldsymbol{t}_m), \tag{9}$$

However, the Euclidean distance is not ideally suitable for complex traffic scenarios with high motion uncertainty. Therefore, in addition to the Euclidean distance, the deep feature representation-based cosine distance between the detections and tracks is produced as the other data association metric. For the detection $\boldsymbol{d}_n$, its appearance feature vector $\boldsymbol{v}_n$ can be generated through the deep CNNs and further normalized to $\|\boldsymbol{v}_n\| = 1$. The smallest feature vector distance is computed as the appearance representation-based metric, given by:

$$\mathscr{M}_{n,m}^{(2)} = min\{1 - \boldsymbol{v}_n^{\mathrm{T}} \boldsymbol{u}_{m,z} \mid \boldsymbol{u}_{m,z} \in \mathscr{U}_m\}, \tag{10}$$

where $\mathscr{U}_m$ represents the set of appearance feature vectors that have been associated for the track $\boldsymbol{t}_m$. Both metrics $\mathscr{M}_{n,m}^{(1)}$ and $\mathscr{M}_{n,m}^{(2)}$ are integrated to form the joint metric $\widetilde{\mathscr{M}}_{n,m}$:

$$\widetilde{\mathscr{M}}_{n,m} = \alpha \mathscr{M}_{n,m}^{(1)} + (1 - \alpha) \mathscr{M}_{n,m}^{(2)}, \tag{11}$$

where α is used to weigh the importance of these two distance measures and set to 0.5. Finally, it is possible to indicate and update the tracking state $\mathscr{S}_n$ for $\boldsymbol{d}_n$, formulated as:

$$\mathscr{S}_n = \begin{cases} tracked, & \exists m, \ \min\limits_m \widetilde{\mathscr{M}}_{n,m} \leqslant \Delta \\ lost, & otherwise \end{cases}, \tag{12}$$

where Δ is defined as the dataset-aware threshold. In practice, the tracklets will be considered as false detections and deleted if they are not successfully associated in the first three consecutive frames. After that, the identity-aware plate streams can be generated in the online matching phase following Algorithm 1. It is worth noting that the developed joint metric introduces robust deep feature representation to compensate for the deficiency of motion information in unconstrained traffic scenarios. With the joint metric, plate tracking essentially announces two intentions, *i.e.* filtering detections to reduce false positives and collecting them into plate streams for tracking identity-based recognition.

3.2.2. Multi-task training

As mentioned above, the well-discriminating appearance representations should be produced by deep neural networks. On the one hand, it has been demonstrated that there are clear mutual benefits of jointly performing detection and tracking in a unified framework [75]. On the other hand, it is reasonable to develop a shared CNN model for the purpose of reducing the computational complexity of deep feature vector generation. Inspired by these facts, a multi-task architecture is proposed to simultaneously perform detection and tracking, as illustrated in Fig. 4, which is end-to-end trainable in the joint formulation via multi-task learning. The multi-task framework is extended on our detection network, where the tracking feature generation branch can be supervised trained by deep cosine metric learning [76].

Given the input video frame I_r, its deep feature maps have been produced by $F_r = N_e(I_r)$. As depicted in Fig. 4, in parallel with three detection substructures, the shallow classification-specific subnetwork N_{cls} is designed for generating deep appearance representation vector $\boldsymbol{v}_n$. With the detection box $\boldsymbol{d}_n$ and global features $F\prime_r$, plate region features can be extracted using *RoIAlign* [77], formulated as $P_{r,n} = RoIAlign(F_r, \boldsymbol{d}_n)$. Moreover, the cosine softmax classifier is exploited at the end of N_{cls} to transform tracking feature generation into the classification task for supervised training, stated by:

$$p(\hat{y}_n = c | \boldsymbol{v}_n) = \frac{exp(\sigma \cdot \boldsymbol{\omega}_c^{\mathrm{T}} \boldsymbol{v}_n)}{\sum\limits_{k=1}^{C} exp(\sigma \cdot \boldsymbol{\omega}_k^{\mathrm{T}} \boldsymbol{v}_n)}, \tag{13}$$

where $\boldsymbol{\omega}$ and σ represent the unit-length weight vector normalized by ℓ_2 and the trainable scaling parameter, respectively. In the above formulation, $\hat{y}_n$ is denoted as the class prediction and $\hat{y}_n, c \in \{1, \cdots, C\}$, where C means the total number of *tracking identities*. Moreover, the joint loss function for detection and classification can be formulated as:

$$\mathscr{L}_{joint} = \mu L_{det} + (1 - \mu) L_{cls} + \eta L_{reg}, \tag{14}$$

where μ represents the fixed weight to balance the detection loss L_{det} and tracking (classification) loss L_{cls}, which is set to 0.8 motivated by [78,79]. L_{reg} is ℓ_2 regularization to alleviate overfitting while η is set to 0.5. In practice, we follow the detection backbone in [40] to design L_{det} while L_{cls} is defined as the softmax cross-entropy loss function, computed by $L_{cls} = -\sum_{n=1}^{N} \log(p(\hat{y}_n = y_n | \boldsymbol{v}_n))$, y_n representing the tracking identity ground truth of $\hat{y}_n$.

As illustrated in Fig. 4, with such multi-task training, the computing resources can be reduced by sharing the feature extraction

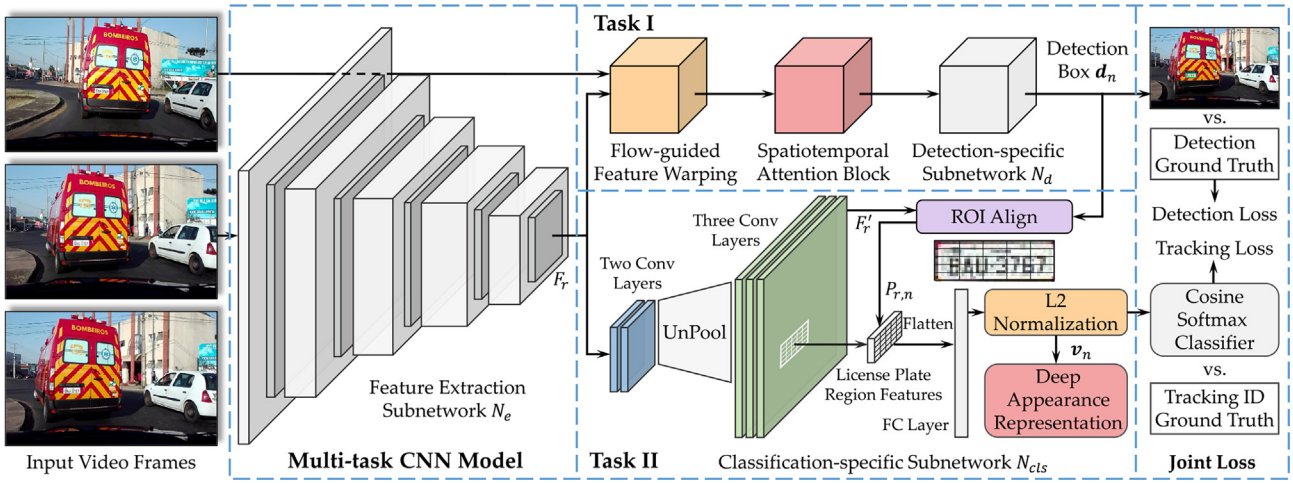

Fig. 4. Multi-task training architecture for license plate detection and tracking.

subnetwork. Furthermore, the well-discriminating deep appearance representations are produced and extracted for online plate tracking during the inference stage.

3.3. Efficient quality-guided license plate stream recognition

Previous methods have introduced majority voting and super-resolution to employ temporal information for plate recognition. However, there are two remarkable issues: (1) In real-world traffic videos, a plate usually continuously appears in its multi-frame stream, where the high-quality frames can be accurately recognized yet the low-quality ones interfere with plate recognition, causing errors. (2) The distortions such as blur, noise, and defocus are commonly global, which significantly affect the image quality of plate patches, whereas this quality is available as prior knowledge.

To address these issues, a plate image quality-guided recognition model is proposed in this section, which consists of two modules, *i.e.* plate recommender and recognizer. We tackle the plate recognition task as a sequence-to-sequence problem. First license plates with the same tracking identity are integrated into one plate stream. Then for each identity-aware stream, the plate recommender is developed to select the highest quality plate online by image quality scoring. Finally, the recommended plate patch is

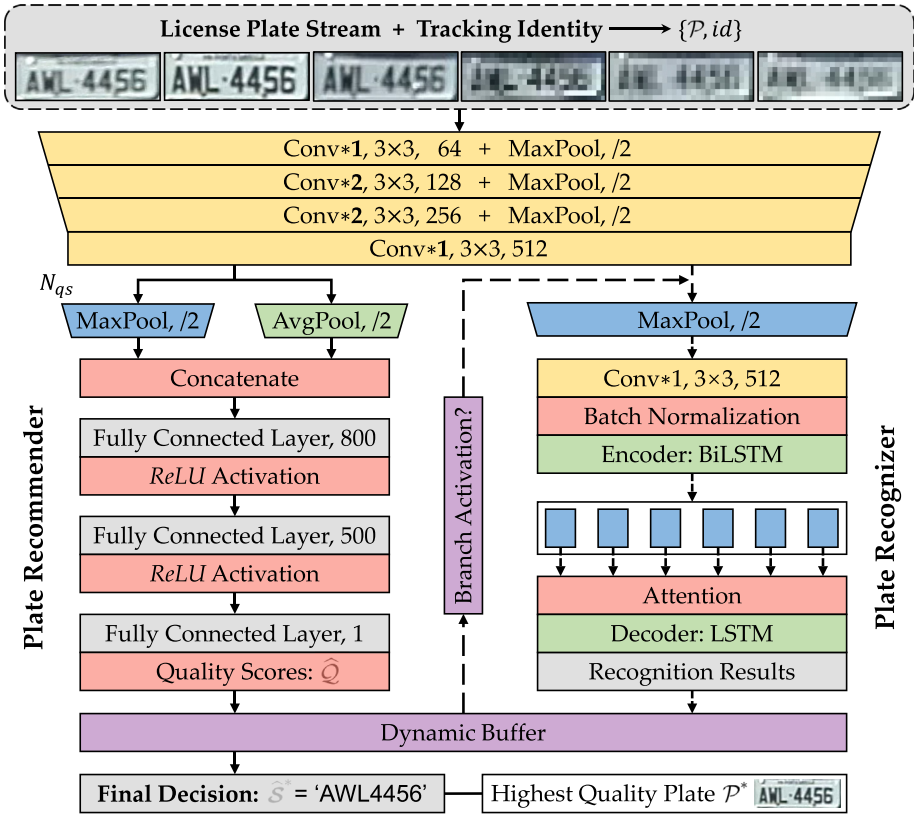

Fig. 5. The proposed multi-task architecture for license plate stream recognition, which is composed of two components: plate recommender and plate recognizer.

fed to the plate recognizer, whose recognition result represents the whole stream's final decision. Note that to accommodate the proposed V-LPDR framework for online applications, there is no future information available in the traffic videos. Against this issue, as shown in Fig. 5, we introduce a dynamic buffer to cache the highest quality score of the current plate stream and its recognition result. In this section, we will introduce the network design and the training strategy in detail.

Algorithm 2. Dynamic Buffer for License Plate Stream Recognition

Input: Quality score $\hat{q}_j \in \widehat{\mathcal{Q}}$ and tracking identity *id*, where *j* represents the serial number of the input license plate $\mathscr{P}_j$ in the stream $\mathscr{P}$

if $j = 1$ **then**
  Initialize:
    Set $\tilde{q}_{id}$ as the highest quality score for the whole plate stream $\mathscr{P}$
    Set $\tilde{\mathscr{S}}_{id,j}$ as the final recognition results of the current input plate $\mathscr{P}_j$
    Activate plate recognizer
      Generate recognition results $\widehat{\mathscr{S}}_1$ of $\mathscr{P}_1$
    Assign $\tilde{q}_{id} \leftarrow \hat{q}_1$, $\tilde{\mathscr{S}}_{id,j} \leftarrow \widehat{\mathscr{S}}_1$
else
  Compare the quality scores $\tilde{q}_{id}$ and $\hat{q}_j$
  if $\hat{q}_j > \tilde{q}_{id}$ **then**
    Update:
      Activate plate recognizer
        Generate recognition results $\widehat{\mathscr{S}}_j$ of $\mathscr{P}_j$
      Assign $\tilde{q}_{id} \leftarrow \hat{q}_j$, $\tilde{\mathscr{S}}_{id,j} \leftarrow \widehat{\mathscr{S}}_j$
    end if
  end if

Output: The final decision $\widehat{\mathscr{S}}^* = \tilde{\mathscr{S}}_{id,j}$ for the input plate $\mathscr{P}_j$

3.3.1. License plate recommender and recognizer

Considering the misrecognition caused by motion blur and perspective distortion, the plate recommender is proposed to process each input plate patch in advance, which involves the quality scoring module and the dynamic buffer, illustrated in Fig. 5. For an identity-aware plate stream, given by $\{\mathscr{P}, id\}$, where $\mathscr{P} = \{\mathscr{P}_1, \mathscr{P}_2, \cdots, \mathscr{P}_J\}$ is denoted as a series of plate patches with the same identity *id*, the quality scoring network N_{qs} is utilized to sequentially evaluate the image quality of each input plate $\mathscr{P}_j \in \mathscr{P}$. In practice, the quality scores predicted by N_{qs} are constrained to the range of $[0, 1]$, formulated as $\widehat{\mathcal{Q}} = N_{qs}(\mathscr{P})$, where $\widehat{\mathcal{Q}} = \{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_J\}$. Then they will be compared based on the input *id* in the buffer. The process of the dynamic buffer initialization and update is summarized in Algorithm 2. After that, for the plate stream $\{\mathscr{P}, id\}$, the recommender outputs the recognition result $\widehat{\mathscr{S}}^*$ corresponding to the highest quality plate $\mathscr{P}^*$ as the final decision.

As presented in Fig. 5, we follow the multi-task design approach to develop the plate recommender and recognizer, where these two tasks share a shallow 6-layer CNN model to extract deep discriminative features. As to quality scoring subnetwork, there are two different pooling operations, max pooling and average pooling, followed by fully connected layers. In line with its input identities and estimated quality scores, whether to activate the recognition branch can be effectively controlled. With the feature maps encoded by the shared CNNs, we exploit the attention-based encoder-decoder model for the plate recognition task, following the previous methods in [80,81]. The recommender performs on every plate patch $\mathscr{P}_j$ yet the recognizer will only be activated under the above specific conditions, contributing to the computational efficiency and temporal redundancy removal.

3.3.2. Weakly supervised alternative training

During the plate recommender and recognizer training phase, there are two different tasks that need to be considered, requiring their corresponding labels. Different from available recognition labels, the quality score ground truths, which indicate the objective quality of plate images, can hardly be manually annotated due to inconsistent subjective assessments and tremendous labor costs. To solve this problem, we innovatively introduce the weakly supervised alternative training strategy based on the notion that higher quality plate patches are more likely to be correctly recognized. In fact, softmax outputs usually imply prior knowledge of image quality [82,83]. Formally, given an input plate patch $\mathscr{P}_j$, its total recognition confidence from softmax function can be computed by:

$$q_j = \prod_{k=1}^{K+2} softmax\left(x_k^{(j)}\right), \tag{15}$$

where $x_k^{(j)}$ denotes the LSTM decoder output for the *k*-th character prediction while $q_j \in \mathcal{Q}$ and $\mathcal{Q} = \{q_1, q_2, \cdots, q_J\}$. K represents the sequence length of the plate $\mathscr{P}_j$ whereas $K+2$ means the extra start and end tokens. q_j will be further used as the ground truth of $\hat{q}_j \in \widehat{\mathcal{Q}}$ for training the quality scoring module, which lies in $[0, 1]$ as well. Therefore, let $\mathscr{S}^{(j)}$ be the recognition ground truth of $\mathscr{P}_j$, and then the loss functions of the recognizer and recommender can be formulated as:

$$L_{rcg} = -\frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K+2} \ln P\left(\mathscr{S}_k^{(j)} | \mathscr{S}_{1:k-1}^{(j)}, \mathscr{P}_j\right), \tag{16}$$

$$L_{qs} = \|\mathcal{Q} - \widehat{\mathcal{Q}}\|_1 = \frac{1}{J}\sum_{j=1}^{J} \|q_j - \hat{q}_j\|, \tag{17}$$

where $\mathscr{S}_k^{(j)}$ denotes the *k*-th location in $\mathscr{S}^{(j)}$. As demonstrated above, since training the plate recommender depends on the softmax prediction of the recognizer, an alternative training strategy is adopted instead of joint training. Briefly speaking, the recognizer is trained at each step yet the recommender will be trained after a certain step interval adjusted as the training progress. It is worth noting that the training of the recommender is weakly supervised with gradually generated weak labels of quality scores rather than exact ones. In practice, two different optimization processes are utilized to minimize the loss functions L_{rcg} and L_{qs} respectively.

4. Experiments

In this section, extensive experiments have been conducted to demonstrate the effectiveness of the proposed framework V-LPDR and evaluate its performance in real-world traffic videos. For clear descriptions, the used license plate datasets and their details are first stated and explained. Then the evaluation criteria and implement settings are briefly introduced. Finally, we present and analyze the experimental results on each dataset for comparison with other competitors.

4.1. Datasets

In order to evaluate the performance of V-LPDR, it is required to access several video-based license plate datasets. However, unlike single image-based datasets in plenty, there are only a few video-based plate datasets available. Among them, three challenging ones are selected and adopted in the experiments, namely UFPR-ALPR dataset [11], SSIG-SegPlate dataset [10], and Low-Quality Plate-Videos dataset [29].

4.1.1. UFPR-ALPR dataset

The first dataset is denoted as the UFPR-ALPR dataset, which is issued by Laroca et al. in [11]. It consists of 150 license plate video clips captured in Brazil, and each video clip is constituted of 30 frames for the same plate. Thus there are a total of 4500 frames in this dataset. Towards the challenging LPDR tasks in unconstrained scenarios, the dataset is built on real-world driving traffic environments with different camera devices, varying illumination, and diverse perspectives. For a closer and detailed inspection of UFPR-ALPR dataset, we derive the statistical map of license plates and their density maps, shown in Fig. 6(a)–(d). In practice, 90 video clips are utilized for training and others for testing. Furthermore, data augmentation by random image cropping and rotation is also implemented during the training phase.

4.1.2. SSIG-SegPlate dataset

The second one is the SSIG-SegPlate dataset released in [10]. It has 2000 license plate frames from 101 video clips of different lengths. With the resolution of 1920×1080 pixels, each plate image in this dataset is acquired by the static digital cameras in real-world traffic scenarios. The statistical map and density maps for this dataset are exhibited in Fig. 6(e)–(h). In the experiments, following the evaluation protocols in [10], 61 video clips are adopted to train our model and others for testing. In addition, data augmentation is implemented to expand the training set as well.

4.1.3. Low-quality plate-videos dataset

The third dataset is proposed in [29] comprising 200 real-world traffic videos (totally 48,000 frames). The movement of all vehicles is away from the video capture device, which results in low-quality video frames with low resolution due to the long distances and fast driving speed. For convenience, we name it "Low-Quality Plate-Videos Dataset" in this paper. Furthermore, it is divided into two subsets, named "Landscape (LA)" and "Portrait (PO)", according to the camera's horizontal or vertical rotated position. It should be noted that the full annotations are NOT provided in [29], while only the recognition labels of the first frame of each video are included in the ground truth files. In order to comprehensively evaluate the proposed method on it, we modify and improve this dataset by manually labeling all visible plates in these 200 videos with both detection and recognition ground truths. Besides, its data distribution is fairly different from the above two datasets, which can benefit the experimental variety. The statistical map and density maps are presented in Fig. 6(i)–(l). In practice, 150 video clips are employed for training and the rest for testing.

4.2. Evaluation measures and implementation setup

To comprehensively evaluate the performance of the proposed unified framework V-LPDR, diverse evaluation measures are exploited for different stages in the algorithm.

4.2.1. Evaluating detection

For evaluating the license plate detection performance, we follow the metrics introduced in [84]. It is reasonable to assume that TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. Conse-

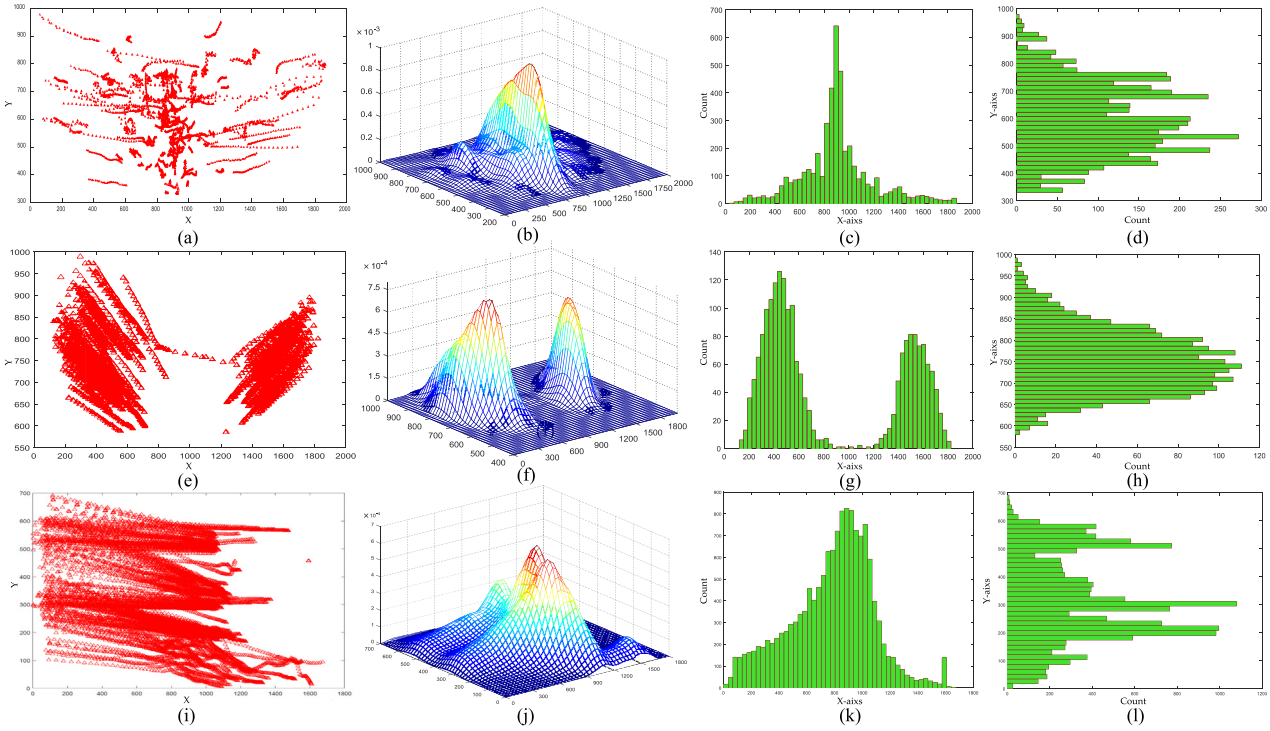

Fig. 6. (a)–(d) for the UFPR-ALPR dataset, (e)–(h) for the SSIG-SegPlate dataset, and (i)–(l) for the Low-Quality Plate-Videos dataset. (a), (e), and (i) represent the statistical maps of the license plate distribution. (b), (f), and (j) mean the probability density maps obtained by 2-dimensional Parzen-window estimation with Gaussian kernel. (c), (g), and (k) are the density maps on X-aixs. (d), (h), and (l) are the density maps on Y-aixs.

quently, the definition of detection recall and precision can be given by:

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP}. \tag{18}$$

Both evaluation protocols are necessary since they serve different purposes for plate detection. Moreover, it is noteworthy that every frame in the testing sets will be evaluated by the above two criteria for the fair comparison. In practice, it will be considered the correct detection if the license plate is encompassed by its bounding box, formulated as $D \cap G/D \cup G \geqslant 0.5$, where D and G represent the localized potential plate region and its corresponding ground truth, respectively.

4.2.2. Evaluating tracking

In order to evaluate the tracking performance, the standard CLEAR MOT metrics defined in [85] are employed in our experiments. The plate tracking criteria should naturally present the tracking results of the proposed plate tracker in real-world traffic videos. Hence we introduce two typical metrics, *i.e.* multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA), to demonstrate its effectiveness.

4.2.3. Evaluating recognition

As to license plate recognition, we evaluate its performance using the sequence recognition accuracy [18], which is defined as the number of correctly recognized license plates divided by the total number of ground truths. It should be reported that accurately recognizing the entire plate sequence (all characters in it) means correct recognition. This metric is quite different from that utilized in some segmentation-based methods [8,11].

In addition, TensorFlow is implemented to build the proposed V-LPDR framework with an Intel Core i7 6800 K @ 3.40 GHz CPU and four NVIDIA GeForce GTX 1080Ti GPUs.

4.3. Ablation studies

To analyze the validity of the proposed modules in V-LPDR, in this section, we conduct some ablation experiments on Low-Quality Plate-Videos dataset. Compared with the other two datasets, this one includes fast-moving vehicles and long distances between the camera and license plates, leading to more serious motion blur and perspective distortion. Thus, this dataset is suitable for ablation studies in the performance improvement brought by the proposed modules like flow-guided spatiotemporal attention and plate stream recommender. Some examples on this dataset are shown in Fig. 7.

4.3.1. Effect of the spatiotemporal detector

To analyze the impact of the proposed spatiotemporal attention mechanism on license plate detection, we conduct ablation experiments using different detection networks. As shown in Table 1, four plate detectors are developed and evaluated on Low-Quality Plate-Videos dataset. "DB" refers to our detection backbone in Section 3, which can be presented only based on single images without any temporal knowledge. The method "DB + Attention" in Table 1 represents the modified "DB" by only introducing the spatial attention, which is also a single-frame detector, while "DB + Optical Flow" includes the process of flow map generation and temporal feature warping but without spatial attention. Naturally, "DB + Spatiotemporal" means the proposed license plate detector, which introduces both temporal optical flow and spatial attention, referred to as "Spatiotemporal Attention" in this paper. The first two are single-image detector, and compared to DB, introducing spatial attention can improve the performance of plate detection, especially the precision, with a maximum increase of 6.11%. This is mainly because it reduces the feature response of false detections. On the other hand, the latter two are video-based detectors, as illustrated in Table 1, in general, they achieve better detection performance due to the benefits from contextual information in optical flow maps. The generation of flow maps does bring extra runtime about 30 ms, but the refined features by optical flow strengthen the response of damaged plate regions with motion blur or partial occlusion in consecutive frames. Therefore, higher detection recall can be obtained, verifying the effectiveness of introducing optical flow for plate detection in videos. The proposed detector ("DB + Spatiotemporal" in Table 1) achieves the best experimental results on the two subsets, and the above ablation study demonstrates the effectiveness of its different modules.

4.3.2. Effect of the online tracker, stream recommender and recognizer

For a general LPDR system, after plate detection, the recognition module will be activated immediately to produce the final decision. In the V-LPDR framework, we perform online plate tracking between detection and recognition since the proposed plate recommender relies on the tracked plate streams to select high-quality frames for the recognizer. In fact, plate tracking is not our ultimate goal, but as a crucial cohesive stage, it will affect the recognition performance. Therefore, in this section, we verify the effect of the proposed tracker and recommender on the recognizer by calculating the recognition accuracy of different models. As shown in Table 2, 5 models are developed and evaluated for ablation studies. We first build single-frame plate recognition algorithms based on the "Only Recognizer" but adopting different model inputs. In Table 2, "by Detection", "by Tracking", and "by Ground Truth" mean that the results of our proposed detector, tracker or ground truths are fed into the recognizer frame by frame as inputs respectively. It can be seen that compared to "by Detection", the recognizer "by Tracking" has higher accuracy with the increases of more than 5%, which proves the benefits from plate tracking mainly due to the detections refined by the tracker. In addition, the combination of recommender and recognizer ("Recommender + Recognizer" in Table 2) is based on tracking, which recommends high-quality license plates from multiple frames to the recognizer and adopts their results instead of low-quality plates. That is why "Recommender + Recognizer" methods can achieve higher recognition accuracy than "Only Recognizer by Ground Truth" and this improvement is significant. Furthermore, we explore the effect of the deep appearance (DA) module in our proposed tracker in the ablation experiments. Specifically, as illustrated in Table 2, we present two models, "Tracking without DA" and "Tracking with DA", where the former means deleting the DA module in the plate tracker while the latter contains it. Both of them are evaluated in the scheme of "Recommender + Recognizer", but "Tracking with DA" obtains the best recognition accuracy at 91.37% and 91.90%. As for computational speed, these two "Recommender + Recognizer" methods perform better since they avoid the plate recognition process on low-quality frames by outputting the results of high-quality frames, and the method "Tracking with DA" has the fastest runtime on this dataset, at only 7 ms per frame. These ablation experiments further verify the effectiveness of our proposed plate tracker, recommender and recognizer.

4.4. Comparison with other methods for each component in V-LPDR

In this section, several experiments including comparative tests are carried out on the UFPR-ALPR dataset to verify the superiority of the proposed framework, referred to as V-LPDR, in unconstrained driving scenarios. We present and analyze the detailed evaluation results for each component in V-LPDR.

Fig. 7. Some plate detection, tracking, and recognition examples on Low-Quality Plate-Videos dataset. The cropped license plate regions are attached to the top left of each frame, followed by their recognition results and tracking identities. The top two rows exhibit the results in LA subset, while the last row presents the results in PO subset. These figures from PO are intercepted for better display.

Table 1
Comparison of different detectors on low-quality plate-videos dataset.

Methods	LA		PO		Speed (ms)
	Recall (%)	Precision (%)	Recall (%)	Precision (%)	
Detection Backbone (DB)	75.19	80.90	79.69	81.76	**85**
DB + Attention	78.46	87.01	81.39	87.22	87
DB + Optical Flow	84.70	85.46	86.53	85.04	118
DB + Spatiotemporal (Optical Flow + Attention)	**90.86**	**94.56**	**91.17**	**94.39**	120

Table 2
Recognition accuracy with different trackers and recognizers on low-quality plate-videos dataset.

Methods		Accuracy (%)		Speed (ms)
		LA	PO	
Only Recognizer	by Detection	65.24	65.67	31
	by Tracking	71.31	70.89	31
	by Ground Truth	87.30	87.48	31
Recommender + Recognizer	Tracking without DA	88.68	88.54	22
	Tracking with DA	**91.37**	**91.90**	**7**

4.4.1. Performance analysis of license plate detection

The UFPR-ALPR dataset has 150 video clips with the bounding box labels of license plates. In this experiment, these annotations are employed to evaluate the performance of the proposed plate detector. The detailed detection results on this dataset are shown in Table 3. Two state-of-the-art methods [11,37] are introduced for detection performance comparison on this dataset. Furthermore, our detection backbone is exploited as the "Baseline"

Table 3
Detection performance comparison on UFPR-ALPR dataset.

Methods	Recall (%)	Precision (%)	No. of true/false detections	Speed (ms)
Laroca et al. [11]	98.33	98.33	1770/30	–
Laroca et al. [37]	98.25	98.67	1776/24	–
Baseline (Ours)	98.50	98.45	1773/28	**85**
Proposed (Ours)	**99.39**	**99.11**	**1789/16**	122

method in Table 3 to verify the benefits of spatiotemporal contextual information for plate detection, especially in complex and unconstrained scenarios. We calculate the detection recall and precision for the comprehensive evaluation, and the proposed plate detector outperforms other algorithms in terms of both evaluation criteria. Concretely, around 1.14% performance improvements of recall can be achieved due to the increased number of true detections. This gain can be mainly attributed to the refined features by temporal information in optical flow maps. At the same time, the detection precision has also been improved by 0.44%, which proves that spatiotemporal attention can reduce the response of false detections and enhance the feature robustness. These experimental results ideally demonstrate the effectiveness and superiority of spatiotemporal attention mechanism for video-based LPDR systems. Moreover, it is satisfactory to obtain such detection performance improvements considering the unconstrained driving scenarios in this dataset. As to the computation speed, the proposed detector takes about 122 ms for each input while the baseline needs 85 ms. The extra cost is worthwhile in terms of performance enhancement. Some detection examples on the UFPR-ALPR dataset are shown in Fig. 8.

4.4.2. Performance analysis of license plate tracking

In this experiment, the performance of the proposed license plate tracker is evaluated on the 60 test video clips from the UFPR-ALPR dataset, the same as used in the detection evaluation. Two typical multiple object tracking metrics, MOTP and MOTA, are exploited in the experiment. It is worth noting that following the strategy in many tracking-by-detection paradigms, the proposed tracker only filters and outputs the bounding boxes generated in the detection phase without additional adjustments. More importantly, it associates all patches of the same license plates with distinguishable identities. Then, these identity-aware plate streams can be efficiently recognized in the next stage. As illustrated in Table 4, we introduce four typical tracking algorithms for comparison. The Kalman filter (KF) and Hungarian algorithm (HA) are implemented in the comparative experiment since the proposed tracker is based on the scheme of "KF + HA", which can also be regarded as our baseline method. About 1.16% and 8.39% performance improvements of MOTP and MOTA are achieved respectively, which demonstrates the effectiveness of introducing deep appearance representation in the joint metric for plate tracking, especially in unconstrained and complex scenarios. We also conduct experiments using two typical MOT methods for comparison, TC_ODAL [86] and MDP [87], both of which outperform "KF + HA". In fact, the tracking performance is evaluated based on the detection results produced by the proposed detector, and as shown in Table 4, four tracking algorithms achieve similar MOTA. This is because the detection performance plays a crucial role in tracking and our detector can produce precise detections for tracking. Both TC_ODAL and MDP are online algorithms but their speed is relatively slow, far from real-time performance. In practice, our well-trained model can generate detection and tracking results simultaneously via the multi-task training strategy introduced in Section 2B. In this way, the runtime of our proposed tracker is reduced to 24 ms, as illustrated in Table 4. This performance improvement especially in accuracy verifies the benefits from the deep appearance module for online plate tracking, and the speed is promising for such complex scenarios. Furthermore, as an online plate tracker, it introduces no future information in videos, which reasonably meets the requirements of real-world applications.

4.4.3. Performance analysis of license plate recognition

Following the evaluation protocol introduced by the dataset administrators in [11], 60 license plate streams (1800 frames) are used to test the performance of the proposed recognition network. For rigorous evaluation, six different methods [88,89,11,90,91,37] are employed for performance comparison in the experiments. Moreover, we also conduct the comparative experiment to demonstrate the benefits of the proposed multi-frame stream recommendation strategy for plate recognition. Specifically, after removing the multi-frame recommendation module, the remaining attention-based single-frame recognition network is utilized as the "Baseline" method in our experiments. For fair comparison and simplicity, the baseline and the proposed approach are trained and tested separately but with the same evaluation protocol. As shown in Table 5, the algorithm developed by Laroca et al. in [37] archived the state-of-the-art recognition performance with about 90% accuracy on the UFPR-ALPR dataset before our method. Nevertheless, their approach may not work well in low-quality license plate images with distortions, leading to a high number of false recognitions. The proposed joint plate recommender and recognizer can address this issue and obtain 95.44% recognition accuracy, which outperforms all other approaches including the baseline. This improvement is mainly because our algorithm avoids performing plate recognition on low-quality frames but adopts the results on high-quality frames instead. The plate recommender can select the current best frame from a tracked plate stream according to their quality scores. Therefore, it achieves both better accuracy and speed. As illustrated in Table 5, it takes about 16 ms per frame, faster than the baseline. This runtime is tolerable and sufficient for some real-world applications in unconstrained scenarios. Overall, the experimental results reveal the effectiveness of quality-guided stream recognition especially in complex environments. Fig. 8 presents several exemplar recognition results on the dataset.

4.5. End-to-end performance comparison with state-of-the-arts

In order to further compare and analyze the end-to-end overall performance of the proposed unified V-LPDR framework with other state-of-the-art algorithms, the experiments are conducted on SSIG-SegPlate dataset, which is more widely used than the other two datasets. In the training phase, 61 video clips from this dataset and all frames in UFPR-ALPR dataset are fed into our model since there are license plates with the same layout in these two datasets. The remaining 40 videos (804 frames) are employed for testing. For rigorous evaluation, both detection and recognition performance are exhibited and compared with state-of-the-art in Table 6.

More importantly, the end-to-end performance and corresponding number of false recognitions are also presented and analyzed in the experiments. To be specific, we introduce nine different competitors for performance comparison. As illustrated

Fig. 8. Some plate detection, tracking, and recognition examples on UFPR-ALPR dataset.

Table 4
License plate tracking performance on UFPR-ALPR dataset.

Methods	MOTP (%)	MOTA (%)	Speed (ms)
Mean-Shift	51.75	42.17	26
KF + HA	93.13	79.72	**18**
TC_ODAL [86]	93.21	81.11	260
MDP [87]	93.74	86.44	400
Proposed tracker	**94.29**	**88.11**	24

in Table 6, their end-to-end accuracy is presented and compared with the proposed method despite the lack of several results for detection or recognition. It can be seen that our approach outperforms all state-of-the-arts and archives the best end-to-end accuracy of 98.64%, which is 1.74% higher than the deep learning-based algorithm in [37]. Accordingly, the total number of end-to-end false recognitions is reduced to 11 on this dataset that has 804 test images. As for the component performance of the proposed V-LPDR, it archives the highest detection recall, precision, and recognition accuracy, respectively 99.88%, 99.63%, and

Table 5
Recognition performance comparison on UFPR-ALPR dataset.

Methods	Recognition accuracy (%)	No. of false recognitions	Speed (ms)
Sighthound [88]	47.39	947	33
OpenALPR [89]	50.94	883	33
Laroca et al. [11]	64.89	632	14
Gonçalves et al. [90]	76.50	–	–
Silva and Jung [91]	85.42	262	20
Laroca et al. [37]	90.00	180	**6**
Baseline (Ours)	92.28	139	32
Proposed (Ours)	**95.44**	**82**	16

Table 6
Performance comparison on SSIG-SegPlate dataset.

Methods	Detection performance (%)		Recognition performance (%)	End-to-End Performance (%)	No. of End-to-End False Recognitions	Overall speed (ms)
	Recall	Precision				
Silva and Jung [92]	99.51	95.07	78.23	63.18	296	115
Sighthound [88]	–	–	82.80	73.13	216	–
Gonçalves et al. [93]	–	–	93.60	81.80	146	250
Laroca et al. [11]	99.75	99.13	92.48	85.45	117	22
OpenALPR [89]	–	–	92.00	87.44	101	–
Silva and Jung [24]	–	–	–	88.56	92	200
Gonçalves et al. [90]	–	–	–	88.80	90	45
Silva and Jung [91]	97.39	96.09	–	92.41	61	115
Laroca et al. [37]	99.78	95.28	98.20	96.90	25	**14**– 34
Ours	**99.88**	**99.63**	**98.88**	**98.64**	**11**	156

Fig. 9. Some exemplar resullts of the proposed method on SSIG-SegPlate dataset.

98.88%. In addition, it also obtains the promising tracking performance of 96.90% MOTP and 89.18% MOTA while there are no quantitative results available from these existing methods that can be collected for comparison. In fact, this end-to-end performance improvement can also be attributed to the effective temporal information in videos introduced by the proposed V-LPDR. To sum up, the above experimental results in Table 6 indicate that the proposed framework has archived the state-of-the-art end-to-end performance on SSIG-SegPlate dataset, which also verifies its validity in traffic videos with slow-moving vehicles and various backgrounds. In addition, some exemplar results on this dataset are exhibited in Fig. 9.

4.6. Discussion

Ablation experiments have verified the effectiveness of each component of the proposed V-LPDR, and the results on different datasets have proven that V-LPDR outperforms other LPDR models and achieves state-of-the-art performance in various scenarios. Moreover, to analyze the proposed V-LPDR architecture more rigorously and comprehensively, we discuss its advantages and disadvantages by comparing it with other recent methods in this section. Different from the existing LPDR algorithms especially deep learning-based models, V-LPDR has four main advantages. (1) In all three stages, *i.e.* license plate detection, tracking and

recognition, V-LPDR takes advantage of contextual information in traffic videos to improve the system accuracy and robustness significantly, while previous models are based on single images, which limits their performance in real-world applications. Experimental results have also demonstrated that exploring the positive temporal knowledge such as optical flow can enhance the LPDR system's performance. (2) As far as we know, V-LPDR is the first work to integrate plate detection, tracking, and recognition into a unified framework via deep learning. The well-trained model can be directly employed for real-world traffic videos in an end-to-end manner without any pre- and post-processing steps, yet many methods only focus on detection or recognition rather than a unified framework. (3) V-LPDR achieves competitive performance in complex and unconstrained scenarios, which is suitable for some intelligent applications such as ITS and Internet of Vehicles. Unlike many LPDR systems for simple scenarios with constrained conditions like parking lots, V-LPDR can obtain improved algorithm robustness when facing some challenges in complex scenarios, for example, multiple vehicles, image/video noise, motion blur, partial occlusion, and varying illumination. (4) The computational cost has been reduced in V-LPDR, especially about the proposed tracker and recognizer. The operation of generating appearance features for tracking is incorporated into the detection network, and the recognizer only processes high-quality plates selected by the recommender instead of recognizing all frames.

However, there are still two disadvantages. (1) Although computation complexity has been reduced by sharing feature extraction and avoiding repeated calculations in the tracking and recognition stage of V-LPDR, its overall processing time for a frame in traffic videos is about $150\,ms/frame$ $(6-7fps)$, which still cannot meet the requirements of some real-time applications. This is mainly attributed to the calculation of optical flow maps in the detection network. (2) As an online algorithm, V-LPDR only exploits the information in past frames from a video without any future knowledge,limiting its performance when the front frames are distorted in some videos. But that is a trade-off between performance and practicality in terms of the LPDR applications for ITS.

5. Conclusion and future work

This work proposes a novel unified framework based on deep learning, namely V-LPDR, for license plate detection, tracking, and recognition in the real-world traffic videos. To be specific, V-LPDR is constituted of three components, spatiotemporal plate detector, online plate tracker, and plate stream recommender and recognizer, which are developed to tackle the above three subtasks, respectively. Firstly, the spatiotemporal detector introduces both temporal information from the past adjacent frames and spatial attention mechanism from the current frame to enhance the plate detection performance especially in unconstrained and complex scenarios. Then, in order to perform plate tracking without future knowledge in videos, a robust online plate tracker is presented based on both motion metrics and deep appearance features, while the latter can be generated and aligned from the detector. Ultimately, the plate recommender pumps out the best quality plate with its tracking identity in the current plate stream, which can be efficiently and accurately recognized by the plate recognizer. To sum up, in this paper, we creatively unify video-based license plate detection, tracking, and recognition into a novel compact and efficient end-to-end framework via deep learning. The proposed approach is evaluated on three different datasets and the ablation experiments and comparison with other methods have verified its effectiveness and robustness in various real-world traffic videos.

Nonetheless, the proposed framework, referred to as V-LPDR, has shortcomings to be tackled in future work. Limited by the computational complexity, it cannot work in real time without performance degradation. As a deep learning-based algorithm, its training and deployment still rely on high-performance computing equipment such as GPUs. In future work, we will introduce model compression like knowledge distillation to V-LPDR to reduce its parameters and improve the processing speed for real-time applications. In addition, deep learning-based LPDR systems including V-LPDR achieve better performance in complex scenarios but lack model interpretability or robust confidence estimation, which limits their adoption in several fields. Thus, explainable LPDR algorithms based on deep learning deserve to be explored in future work.

CRediT authorship contribution statement

Cong Zhang: Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Qi Wang:** Supervision, Writing - original draft, Writing - review & editing. **Xuelong Li:** Writing - original draft, Writing - review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, and 61871470.

References

[1] Q. Wang, J. Gao, Y. Yuan, Embedding structured contour and location prior in siamesed fully convolutional networks for road detection, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 230–241.
[2] Q. Wang, M. Chen, F. Nie, X. Li, Detecting coherent groups in crowd scenes by multiview clustering, IEEE Trans. Pattern Anal. Mach. Intell. 42 (1) (2020) 46–58.
[3] Y. Yuan, Z. Xiong, Q. Wang, An incremental framework for video-based traffic sign detection, tracking, and recognition, IEEE Trans. Intell. Transp. Syst. 18 (7) (2017) 1918–1929.
[4] L. Xie, T. Ahmad, L. Jin, Y. Liu, S. Zhang, A new cnn-based method for multi-directional car license plate detection, IEEE Trans. Intell. Transp. Syst. 19 (2) (2018) 507–517.
[5] S. Du, M. Ibrahim, M. Shehata, W. Badawy, Automatic license plate recognition (alpr): a state-of-the-art review, IEEE Trans. Circuits Syst. Video Technol. 23 (2) (2013) 311–325.
[6] Caltech plate dataset [Online]. Available: http://www.vision.caltech. edu/html-files/archive.html (2003)..
[7] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, N. Komodakis, A robust and efficient approach to license plate detection, IEEE Trans. Image Process. 26 (3) (2017) 1102–1114.
[8] G. Hsu, J. Chen, Y. Chung, Application-oriented license plate recognition, IEEE Trans. Veh. Technol. 62 (2) (2013) 552–561.
[9] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, L. Huang, Towards end-to-end license plate detection and recognition: a large dataset and baseline, European Conference on Computer Vision (2018) 255–271.
[10] G.R. Gonçalves, S.P.G. da Silva, D. Menotti, W.R. Schwartz, Benchmark for license plate character segmentation, J. Electron. Imag. 25 (5) (2016) 053034.
[11] R. Laroca, E. Severo, L.A. Zanlorensi, L.S. Oliveira, G.R. Gonçalves, W.R. Schwartz, D. Menotti, A robust real-time automatic license plate recognition based on the yolo detector, in: International Joint Conference on Neural Networks, 2018, pp. 1–10..
[12] L. Zhang, P. Wang, H. Li, Z. Li, C. Shen, Y. Zhang, A robust attentional framework for license plate recognition in the wild, IEEE Trans. Intell. Transp. Syst. (2020), https://doi.org/10.1109/TITS.2020.3000072.
[13] C.-N.E. Anagnostopoulos, I.E. Anagnostopoulos, I.D. Psoroulas, V. Loumos, E. Kayafas, License plate recognition from still images and video sequences: a survey, IEEE Trans. Intell. Transp. Syst. 9 (3) (2008) 377–391.

[14] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (2014) 580–587.
[15] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, IEEE Conference on Computer Vision and Pattern Recognition (2016) 779–788.
[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: ttowards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 6 (2017) 1137–1149.
[17] J. Zhang, S. Ding, N. Zhang, An overview on probability undirected graphs and their applications in image processing, Neurocomputing 321 (2018) 156–168.
[18] H. Li, P. Wang, M. You, C. Shen, Reading car license plates using deep neural networks, Image Vis. Comput. 72 (2018) 14–23.
[19] H. Li, P. Wang, C. Shen, Toward end-to-end car license plate detection and recognition with deep neural networks, IEEE Trans. Intell. Transp. Syst. 20 (3) (2019) 1126–1136.
[20] Q. Huang, Z. Cai, T. Lan, A new approach for character recognition of multi-style vehicle license plates, IEEE Trans. Multimedia (2020), https://doi.org/10.1109/TMM.2020.3031074.
[21] C. Liu, F. Chang, Hybrid cascade structure for license plate detection in large visual surveillance scenes, IEEE Trans. Intell. Transp. Syst. 20 (6) (2019) 2122–2135.
[22] S. Chen, C. Yang, J. Ma, F. Chen, X. Yin, Simultaneous end-to-end vehicle and license plate detection with multi-branch attention neural network, IEEE Trans. Intell. Transp. Syst. 21 (9) (2020) 3686–3695.
[23] C.P. Chen, B. Wang, Random-positioned license plate recognition using hybrid broad learning system and convolutional networks, IEEE Trans. Intell. Transp. Syst. (2020), https://doi.org/10.1109/TITS.2020.3011937.
[24] S.M. Silva, C.R. Jung, License plate detection and recognition in unconstrained scenarios, European Conference on Computer Vision (2018) 593–609.
[25] M.R. Asif, C. Qi, T. Wang, M.S. Fareed, S.A. Raza, License plate detection for multi-national vehicles: an illumination invariant approach in multi-lane environment, Comput. Electr. Eng. 78 (2019) 132–147.
[26] W. Wang, J. Tu, Research on license plate recognition algorithms based on deep learning in complex environment, IEEE Access 8 (2020) 91661–91675.
[27] M.S. Al-Shemarry, Y. Li, S. Abdulla, An efficient texture descriptor for the detection of license plates from vehicle images in difficult conditions, IEEE Trans. Intell. Transp. Syst. 21 (2) (2020) 553–564.
[28] G.R. Gonçalves, D. Menotti, W.R. Schwartz, License plate recognition based on temporal redundancy, IEEE International Conference on Intelligent Transportation Systems (2016) 1–5.
[29] H. Seibel, S. Goldenstein, A. Rocha, Eyes on the target: super-resolution and license-plate recognition in low-quality surveillance videos, IEEE Access 5 (2017) 20020–20035.
[30] V. Vašek, V. Franc, M. Urban, License plate recognition and super-resolution from low-resolution videos by convolutional neural networks, British Machine Vision Conference (2018).
[31] N. Thome, A. Vacavant, L. Robinault, S. Miguet, A cognitive and video-based approach for multinational license plate recognition, Mach. Vis. Appl. 22 (2) (2011) 389–407.
[32] D. Zheng, Y. Zhao, J. Wang, An efficient method of license plate location, Pattern Recogn. Lett. 26 (15) (2005) 2431–2438.
[33] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, N. Komodakis, A robust and efficient approach to license plate detection, IEEE Trans. Image Process. 26 (3) (2017) 1102–1114.
[34] A.H. Ashtari, M.J. Nordin, M. Fathy, An iranian license plate recognition system based on color features, IEEE Trans. Intell. Transp. Syst. 15 (4) (2014) 1690–1705.
[35] S. Yu, B. Li, Q. Zhang, C. Liu, M.Q.-H. Meng, A novel license plate location method based on wavelet transform and emd analysis, Pattern Recogn. 48 (1) (2015) 114–125.
[36] M. Molina-Moreno, I. González-Díaz, F. Díaz-de María, Efficient scale-adaptive license plate detection system, IEEE Trans. Intell. Transp. Syst. 20 (6) (2019) 2109–2121.
[37] R. Laroca, L.A. Zanlorensi, G.R. Gonçalves, E. Todt, W.R. Schwartz, D. Menotti, An efficient and layout-independent automatic license plate recognition system based on the yolo detector, arXiv preprint arXiv:1909.01754..
[38] A. Tourani, A. Shahbahrami, S. Soroori, S. Khazaee, C.Y. Suen, A robust deep learning approach for automatic iranian vehicle license plate detection and recognition for surveillance systems, IEEE Access 8 (2020) 201317–201330.
[39] C. Henry, S.Y. Ahn, S.-W. Lee, Multinational license plate recognition using generalized character sequence detection, IEEE Access 8 (2020) 35185–35199.
[40] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, East: an efficient and accurate scene text detector, IEEE Conference on Computer Vision and Pattern Recognition (2017) 5551–5560.
[41] Z. Cheng, J. Lu, Y. Niu, S. Pu, F. Wu, S. Zhou, You only recognize once: towards fast video text spotting, ACM International Conference on Multimedia (2019) 855–863.
[42] G. Hsu, J. Chen, Y. Chung, Application-oriented license plate recognition, IEEE Trans. Veh. Technol. 62 (2) (2013) 552–561.
[43] I.V. Pustokhina, D.A. Pustokhin, J.J. Rodrigues, D. Gupta, A. Khanna, K. Shankar, C. Seo, G.P. Joshi, Automatic vehicle license plate recognition using optimal k-means with convolutional neural network for intelligent transportation systems, IEEE Access 8 (2020) 92907–92917.
[44] S. Rasheed, A. Naeem, O. Ishaq, Automated number plate recognition using hough lines and template matching, in: World Congress on Engineering and Computer Science, 2012, pp. 24–26..
[45] J. Jiao, Q. Ye, Q. Huang, A configurable method for multi-style license plate recognition, Pattern Recogn. 42 (3) (2009) 358–369.
[46] I. Giannoukos, C.-N. Anagnostopoulos, V. Loumos, E. Kayafas, Operator context scanning to support high segmentation rates for real time license plate recognition, Pattern Recogn. 43 (11) (2010) 3866–3878.
[47] Y. Wen, Y. Lu, J. Yan, Z. Zhou, K.M. von Deneen, P. Shi, An algorithm for license plate recognition applied to intelligent transportation system, IEEE Trans. Intell. Transp. Syst. 12 (3) (2011) 830–845.
[48] J. Tang, C. Deng, G.-B. Huang, Extreme learning machine for multilayer perceptron, IEEE Trans. Neural Networks Learn. Syst. 27 (4) (2015) 809–821.
[49] N. Zhang, S. Ding, Unsupervised and semi-supervised extreme learning machine with wavelet kernel for high dimensional data, Memetic Comput. 9 (2) (2017) 129–139.
[50] J. Zhang, S. Ding, N. Zhang, Z. Shi, Incremental extreme learning machine based on deep feature embedded, Int. J. Mach. Learn. Cybern. 7 (1) (2016) 111–120.
[51] Y. Yang, D. Li, Z. Duan, Chinese vehicle license plate recognition using kernel-based extreme learning machine with deep convolutional features, IET Intel. Transport Syst. 12 (3) (2017) 213–219.
[52] Z. Huang, H. Tseng, C. Chen, Application of extreme learning machine to automatic license plate recognition, IEEE Conference on Industrial Electronics and Applications (2019) 1447–1452.
[53] O. Bulan, V. Kozitsky, P. Ramesh, M. Shreve, Segmentation-and annotation-free license plate recognition with deep localization and failure identification, IEEE Trans. Intell. Transp. Syst. 18 (9) (2017) 2351–2363.
[54] M. Dong, D. He, C. Luo, D. Liu, W. Zeng, A cnn-based approach for automatic license plate recognition in the wild., British Machine Vision Conference (2017).
[55] G.J. Hsu, C. Chiu, A comparison study on real-time tracking motorcycle license plates, in: IEEE Image, Video, and Multidimensional Signal Processing Workshop, 2016, pp. 1–5..
[56] Y. Yanamura, M. Goto, D. Nishiyama, M. Soga, H. Nakatani, H. Saji, Extraction and tracking of the license plate using hough transform and voted block matching, IEEE Intelligent Vehicles Symposium (2003) 243–246.
[57] I.K. Yalcin, M. Gokmen, Integrating differential evolution and condensation algorithms for license plate tracking, in: IEEE International Conference on Pattern Recognition, 2006, pp. 658–661..
[58] M. Zayed, J. Boonaert, M. Bayart, License plate tracking for car following with a single camera, IEEE Intelligent Transportation Systems Conference (2004) 719–724.
[59] M. Donoser, C. Arth, H. Bischof, Detecting, tracking and recognizing license plates, Asian Conference on Computer Vision (2007) 447–456.
[60] D.C. Luvizon, B.T. Nassu, R. Minetto, Vehicle speed estimation by license plate detection and tracking, IEEE International Conference on Acoustics, Speech and Signal Processing (2014) 6563–6567.
[61] D.C. Luvizon, B.T. Nassu, R. Minetto, A video-based system for vehicle speed measurement in urban roadways, IEEE Trans. Intell. Transp. Syst. 18 (6) (2017) 1393–1404.
[62] M. Zhang, W. Liu, H. Ma, Joint license plate super-resolution and recognition in one multi-task gan framework, IEEE International Conference on Acoustics, Speech and Signal Processing (2018) 1443–1447.
[63] W. Liu, X. Liu, H. Ma, P. Cheng, Beyond human-level license plate super-resolution with progressive vehicle search and domain priori gan, ACM International Conference on Multimedia (2017) 1618–1626.
[64] Y. Zou, Y. Wang, W. Guan, W. Wang, Semantic super-resolution for extremely low-resolution vehicle license plate, IEEE International Conference on Acoustics, Speech and Signal Processing (2019) 3772–3776.
[65] X. Zhu, Y. Wang, J. Dai, L. Yuan, Y. Wei, Flow-guided feature aggregation for video object detection, IEEE International Conference on Computer Vision (2017) 408–417.
[66] X. Zhu, Y. Xiong, J. Dai, L. Yuan, Y. Wei, Deep feature flow for video recognition, IEEE Conference on Computer Vision and Pattern Recognition (2017) 2349–2358.
[67] X. Zhu, J. Dai, L. Yuan, Y. Wei, Towards high performance video object detection, IEEE Conference on Computer Vision and Pattern Recognition (2018) 7210–7218.
[68] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.
[69] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: learning optical flow with convolutional networks, IEEE International Conference on Computer Vision (2015) 2758–2766.
[70] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning, IEEE Conference on Computer Vision and Pattern Recognition (2017) 5659–5667.
[71] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: convolutional block attention module, European Conference on Computer Vision (2018) 3–19.
[72] J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, Bam: bottleneck attention module, British Machine Vision Conferenc (2018) 147.
[73] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, J. Yan, POI: multiple object tracking with high performance detection and appearance feature, European Conference on Computer Vision (2016) 36–42.

[74] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, IEEE International Conference on Image Processing (2017) 3645–3649.
[75] C. Feichtenhofer, A. Pinz, A. Zisserman, Detect to track and track to detect, IEEE International Conference on Computer Vision (2017) 3038–3046.
[76] N. Wojke, A. Bewley, Deep cosine metric learning for person re-identification, IEEE Winter Conference on Applications of Computer Vision (2018) 748–756.
[77] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, IEEE International Conference on Computer Vision (2017) 2961–2969.
[78] O. Sener, V. Koltun, Multi-task learning as multi-objective optimization, Adv. Neural Inf. Process. Syst. 31 (2018) 527–538.
[79] M. Guo, A. Haque, D.-A. Huang, S. Yeung, L. Fei-Fei, Dynamic task prioritization for multitask learning, European Conference on Computer Vision (2018) 270–287.
[80] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, Aster: an attentional scene text recognizer with flexible rectification, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2018) 2035–2048.
[81] Y. Deng, A. Kanervisto, J. Ling, A.M. Rush, Image-to-markup generation with coarse-to-fine attention, International Conference on Machine Learning 70 (2017) 980–989.
[82] A. Subramanya, S. Srinivas, R.V. Babu, Confidence estimation in deep neural networks via density modelling, arXiv preprint arXiv:1707.07013..
[83] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014..
[84] W. Zhou, H. Li, Y. Lu, Q. Tian, Principal visual word discovery for automatic license plate detection, IEEE Trans. Image Process. 21 (9) (2012) 4269–4279.
[85] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear mot metrics, J. Image Video Process. 2008 (2008) 1.
[86] S.-H. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, IEEE Conference on Computer Vision and Pattern Recognition (2014) 1218–1225.
[87] Y. Xiang, A. Alahi, S. Savarese, Learning to track: online multi-object tracking by decision making, IEEE International Conference on Computer Vision (2015) 4705–4713.
[88] S.Z. Masood, G. Shu, A. Dehghan, E.G. Ortiz, License plate detection and recognition using deeply learned convolutional neural networks, arXiv preprint arXiv:1703.07330..
[89] OpenALPR Cloud API, [Online]. Available: http://www.openalpr.com/ cloud-api.html..
[90] G.R. Gonçalves, M.A. Diniz, R. Laroca, D. Menotti, W.R. Schwartz, Real-time automatic license plate recognition through deep multi-task networks, IEEE Conference on Graphics, Patterns and Images (2018) 110–117.
[91] S.M. Silva, C.R. Jung, Real-time license plate detection and recognition using deep convolutional neural networks, J. Vis. Commun. Image Represent. 102773 (2020).
[92] S.M. Silva, C.R. Jung, Real-time brazilian license plate detection and recognition using deep convolutional neural networks, SIBGRAPI Conference on Graphics, Patterns and Images (2017) 55–62.
[93] G.R. Gonçalves, D. Menotti, W.R. Schwartz, License plate recognition based on temporal redundancy, IEEE International Conference on Intelligent Transportation Systems (2016) 2577–2582.

Cong Zhang received the B.E. degree in Northwestern Polytechnical University and is currently pursuing the M.E. degree with the School of Computer Science and Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research interest include machine learning and pattern recognition.

Qi Wang received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Xuelong Li is a Full Professor with the School of Computer Science and Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China.