# An End-to-End Contrastive License Plate Detector

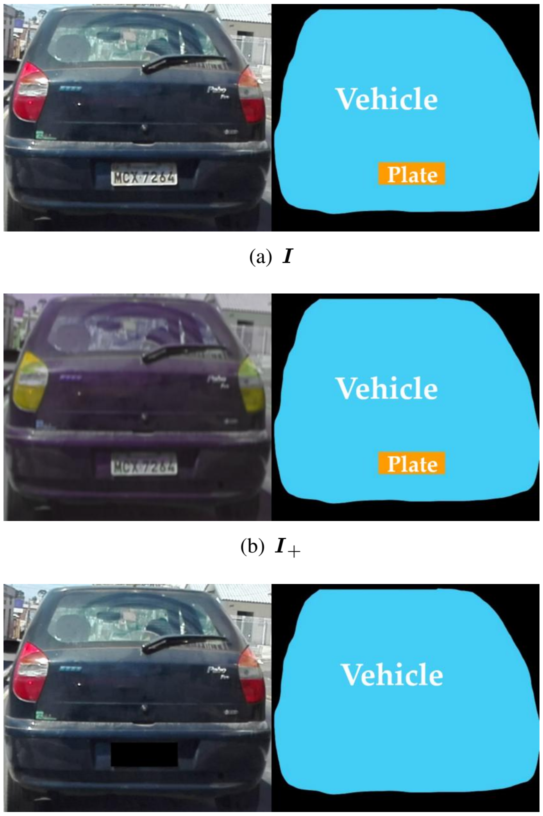
Haoxuan Ding, Junyu Gao, *Member, IEEE,* Yuan Yuan, *Senior Member, IEEE,*
and Qi Wang, *Senior Member, IEEE*

***Abstract*—As a unique identity of vehicle, License Plate (LP) facilitates the intelligent transportation in many fields, such as traffic enforcement, intelligent transportation dispatching, *etc.* Recently, the LP detectors are trained by supervised learning which is directly guided by manual annotations and lacks the use of visual knowledge in image content, limiting the further development of detection performance. Inspired by the contrast and comparison in perception of human beings, a contrastive learning method is introduced into license plate detection task and we propose an end-to-end Contrastive License Plate Detector (CLPD). In CLPD, a special contrastive triad for contrastive learning is designed which aims to decouple the foregrounds and backgrounds. Based on this triad, a contrastive learning branch is introduced into the license plate detection pipeline to prompt the feature expression ability of backbone and extracting more discriminative features for detection. This contrastive learning branch is jointly trained with supervised learning branch for detection and it is only used in training, keeping the efficiency in inference. The experiment results show that the proposed CLPD improves the detection accuracy compared to baselines and other license plate detectors significantly on three datasets. The ablation studies further explore the potential of CLPD. In addition, the proposed CLPD has generalization to improve the performance on different baselines. And the visualization results in latent space verify our proposed CLPD aggregates features tightly and extracts discriminative features effectively.**

***Index Terms*—Automatic license plate detection, Self-supervised learning, Contrastive learning, Feature aggregation.**

## I. Introduction

The Automatic License Plate Recognition (ALPR) systems [1]–[4] have achieved successful application in intelligent transportation and smart city. As a fundamental issue in transportation system, the detection and recognition of license plates (LPs) facilitate many fields in traffic scenes, such as traffic enforcement, transportation dispatching, *etc.* The popular pipeline of ALPR system includes License Plate Detection (LPD) module and License Plate Recognition (LPR) module. The image is first input to LPD for LPs localization, and then the patches of LPs are fed to LPR for recognition of LP numbers. In this cascade pipeline, the accurate detection of LPs has a crucial influence on downstream LP recognition performance. Thus, a high-performance LP detector is necessary in ALPR system.



(a) $\boldsymbol{I}$

(b) $\boldsymbol{I}_+$

(c) $\boldsymbol{I}_-$

Fig. 1. Human beings can easily find the similar and dissimilar instances in image by comparison and contrast.

With the feature extraction ability of Neural Network (NN), recent detectors based on deep learning learn object-related features by supervised learning. However, traditional supervised learning only minimizes the error between predictions and ground-truths without any guidance from prior visual knowledge, limiting the further improvement of detection accuracy. On the contrary, in the visual perception of humans, the foreground are learned through comparison and contrast from backgrounds, prompting the perception accuracy. For example, given original image Fig. 1(a), augmented image Fig. 1(b), and image without LPs Fig. 1(c), human beings easily find the LP in Fig. 1(b), even the degeneration occurs in Fig. 1(b). The contrast of contents in these images provides prior knowledge which helps human beings in decoupling foregrounds and backgrounds.

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902.

H. Ding is with the Unmanned System Research Institute, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P. R. China. (e-mail: haoxuan.ding@mail.nwpu.edu.cn)

J. Gao, Y. Yuan, and Q. Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: gjy3035@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com)

Q. Wang and J. Gao are the corresponding authors.

According to visual perception of human, prior visual knowledge prompts the detection performance significantly. In this context, contrastive learning is proposed to imitate the contrast in visual perception of human. Contrastive learning aims to mine the similarity and dissimilarity in training samples. In this paper, a one-stage LP detector with contrastive learning, named Contrastive License Plate Detector (CLPD), is proposed. We design a contrastive strategy to deeply exploit the visual knowledge as clues to decouple the foregrounds and backgrounds for LP detection task. First, a special contrastive triad is build for contrastive learning. As demonstrated in Fig. 1, this triad contains original image $\boldsymbol{I}$, negative image $I_-$ generated by covering the LPs, and positive image $I_+$ generated by a degeneration strategy. Second, a sharing backbone extracts the latent features from items in contrastive triad, and mines the specificity of LPs via a contrastive learning strategy based on InfoNCE loss [5]. Third, the contrastive learning is jointly trained with supervised learning for detection task.

The backbone learns the dissimilarity between foregrounds and backgrounds through contrastive learning, improving the backbone to extract more discriminative features for detection task. The discriminative features are concentrated on LPs tightly, making higher detection accuracy. The experiments results prove that the proposed CLPD achieves notable improvement compared to baseline and other LPD methods. And we explore the change on latent feature space when contrastive learning is introduced via visualization.

In summary, the main contribution of this paper are:

1) Introduce the contrative learning into LP detection task and propose Contrastive License Plate Detector (CLPD). The traditional supervised learning for detection lacks of the guidance from abundant visual knowledge in images. Thus, to fully exploit the information from vision, we propose a contrative strategy to make the detector decouple the LPs from backgrounds, improving the detection accuracy effectively. The contrative learning branch is trained jointly with supervised learning branch for detection to aggregate foreground features.
2) To explain and analyze the reason why the joint training of contrastive learning is effective in LP detection, we visualize the changes in latent feature space. From the visualization results, the effect of proposed CLPD is verified credibly. The joint training of contrastive learning in CLPD successfully decouples the foregrounds and backgrounds which alleviates the disturbances from backgrounds in LP detection task.
3) The experiment results show that the proposed CLPD exceeds the baseline and other LP detectors significantly. And the contrastive learning branch in CLPD also suits for other baselines, which indicates the contrastive strategy in proposed CLPD has the generalization ability among different detectors and has the potential to suit for other detection tasks effectively.

The rest of this paper is organized as follows. Section II reviews the related work about detection and contrastive learning. Section III describes the details of proposed CLPD. Section IV shows the experiment results. And Section Vdiscusses some issues about the effectiveness of proposed CLPD. Finally, Section VI summarizes the whole work.

## II. Related Work

In this section, we briefly review the related researches about LP detection task. The previous works about general object detection are introduced in Section II-A. Then we summarize the recent works about LP detection in Section II-B. Finally, the recent works about contrastive learning are illustrated in Section II-C.

### A. Object Detection

The general object detection methods mainly contains two categories: two-stage detectors and one-stage detectors. R-CNN [6] is the first two-stage detectors. Along with the pipeline of R-CNN [6], Fast R-CNN [7], Faster R-CNN [8], and Mask R-CNN [9] are proposed. The two-stage detectors first extract a series of region proposals including objects or backgrounds. Then, these region proposals are standardized into fixed-size feature maps by RoIPooling layer [8] or RoIAlign layer [9]. Next, the object categories and object bounding boxes are predicted from feature maps by detection heads. The region proposals first localize the objects roughly, and then the detection heads analyze the information in region proposals for classification and localization. Thus, two-stage detectors achieve accurate detection performance. However, the generation of region proposals is time-consuming and inefficient, causing slow inference speed.

To increase the efficiency, the researchers attempt to remove the region proposals and directly predict the objects categories and positions, proposing the one-stage detectors. The one-stage detectors mainly contain YOLO [10] and its variants YOLOv2 [11], YOLOv3 [12], YOLOv4 [13] *etc.* YOLO-based methods are prevalent one-stage detectors and have been applied in the industrial fields successfully. In addition, some methods break the traditional detection pipeline. CornerNet [14] and CenterNet [15] predict the corner points or center points to localize the objects. FCOS [16] introduces the Fully Convolutional Network (FCN) [17] into object detection task. The one-stage detectors achieve higher efficiency compared to two-stage detectors, making the industrial application of object detection possible. However, one-stage detectors suffer from the imbalance of objects and backgrounds due to lack of region proposals, causing lower accuracy than two-stage detectors. RetinaNet [18] uses focal loss to alleviate the imbalance of foregrounds and backgrounds.

Recently, the transformer [19] has introduced into computer vision. DETR [20] is the first work to introduce vision transformer into object detection task. DETR [20] is a pure end-to-end framework without any hand-made operations, treating the object detection task as a set prediction task. DeformDETR [21] uses deformable attention to increase the convergence speed. Gao *et al.* [22] design a Spatially Modulated Co-Attention (SMCA) mechanism which introduces a location-aware co-attention to concentrate on local features near the estimated bounding boxes. CondDETR [23] introduces a conditional spatial query for box regression and accelerates the training. However, Wang *et al.* [24] finds that the detection transformer only achieves superior performance under large training data (*i.e.* MS COCO dataset [25]) and the performance decreases rapidly when the dataset is small. The LP detection datasets [26], [27] are small and this is the reason why we do not use transformer framework in LP detection task.

### *B. LP Detection*

As a sub-task of object detection, recent LP detectors are mainly inherited from object detection methods introduced above. Two-stage detectors [28]–[30] are first used in LP detection task. To suit for the real-time task, the majority of LP detectors choose one-stage detection frameworks, such as YOLO [10] and its variants [11], [12], [12], [13]. Hsu *et al.* [31] use YOLO [10] and YOLOv2 [11] in LP detection task. Laroca *et al.* [26] build a LP detection dataset named UFPR-ALPR, and then use YOLO [10] and Fast-YOLO [11] in LP detection task. Furthermore, Laroca *et al.* [32] consider the different layout of LPs in detection. Kong *et al.* [33] design a fast LP detection model inspired by YOLO [10] and Mask R-CNN [9], and this model is applied on mobile devices. Jamtsho *et al.* [34] utilize YOLO [10] to detect the LPs of non-helmeted motorcyclists. Zandi *et al.* [35] use YOLOv3 [12] to detect the Iranian LPs.

In addition, many recent works focus on the methods to improve the accuracy and robustness of LP detectors. Silva and Jung [36] propose the Warped Planar Object Detection Network (WPOD-NET) to detect LPs in unconstrained scenarios. Chen *et al.* [37] utilize an end-to-end framework to detect LPs and vehicles simultaneously, and a multi-branch attention mechanism is designed to improve performance. Al-Shemarry *et al.* [4] proposes an multi-level extended local binary pattern descriptor for LP detection. Lee *et al.* [38] design an end-to-end LP detection framework use a shared backbone to detect LPs and scene text simultaneously, distinguishing the difference between LPs and scene texts. Chen *et al.* [39] use the vehicle-plate relation to localize degraded LPs. Silva and Jung [40] present an Improved Warped Planar Object Detection Network (IWPOD-NET) to detect the four corners of an LP in a variety of conditions. Fan and Zhao [41] propose CA-CenterNet to detect the center and four corner points of LPs. Chen and Wang [42] design a neural network with five fully convolutional blocks and five deconvolutional layers to segment the LP masks. Wang *et al.* [43] propose a VertexNet to detect four corner points of LPs. Pirgazi *et al.* [44] use SSD [45] to detect vehicle first and then use a designed CNN to detect LPs.

As for video LP detection task, the temporal relation is introduced into detection. Zang *et al.* [46] design a deep flow-guided spatiotemporal license plate detector to model the video contextual information in LP detection task. Meanwhile, Lu *et al.* [47] also integrate optical flow extraction module in LP detector, which propagates the features of local frames and fuse with the reference frame. Wang *et al.* [48] propose a large-scale video LP detection dataset.

As for contrastive learning for LP detection task, recent LP detectors are lack of the guidance from contrastive learning. And only Lee *et al.* [38] attempt to compare the LPs and non-LPs via a two-stream detector. Thus, recent LP detection works with contrastive learning are still extremely rare.

### *C. Contrastive Learning*

Contrastive learning is an emerging conception in deep learning. It concentrates on the similarity and dissimilarity among data to aggregate the data from same categories together. Inspired by the contrast and comparison in human perception and inference, the contrastive learning makes the representations from similar instances closer and pushes the representations from dissimilar instances far away. Wu *et al.* [49] find that the visual correlated instances have similar outputs from classifier and use this clue to train an unsupervised feature learning approach via instance-level discrimination. Ye *et al.* [50] propose an end-to-end unsupervised learning framework with data augmentation in mini-batch. In this framework, image features and its corresponding augmentation features should be invariant, while the features from other images in the mini-batch should be spread out.

CPC [5] first introduces the contrastive learning into generation task to predict the future words in speech and proposes a well-known contrastive loss: InfoNCE loss. CMC [51] focuses on multi-view contrastive learning. The views from a same instance are similar and they are different from that of another instances. MoCo [52] defines the contrastive learning as a dictionary look-up task and uses dictionary queue and momentum encoder to ensure the diversity and consistency of contrastive keys, achieving promising performance close to supervised learning. SimCLR [53] introduces diverse augmentation methods to build contrastive samples and large size mini-batch. Meanwhile, SimCLR [53] designs a feature projector to improve the performance effectively. Inspired by SimCLR [53], MoCo-v2 [54] introduces data agumentations, projector, and cosine learning rate schedule into MoCo [52]. To abstract the contrastive samples, SwAV [55] first clusters the sample representations and contrasts the query representations with cluster centers (*i.e.* prototype). In special vision tasks, contrastive learning also improves the performance significantly. For example, Hsu *et al* [56] use a triplet loss to compute the similarities and differences from vehicle instances within a batch, enhancing the accuracy and robustness of vehicle re-identification (Re-ID). Gao *et al.* [57] compare the similarity between the coarse map and the Gaussian kernel to improve the counting performance. Wang *et al.* [58] utilize a triplet loss in object tracking task to distinguish different objects in tracklets. In addition, Wang *et al.* [59] also introduce triplet loss into a graph-based method for tracking to better associate the box embedding in graph. In above works, the different contrastive strategies achieve different effects and we could design special contrastive strategy to fulfill our requirements in different tasks.

The above methods all contrast the original samples with both of positive samples and negative samples, BYOL [60] only uses the originals and positives to achieve the contrastive learning without negatives. SimSiam [61] also discards the negative samples. It uses a Siamese Network with a predictor to make every encoder predict the output of another encoder.

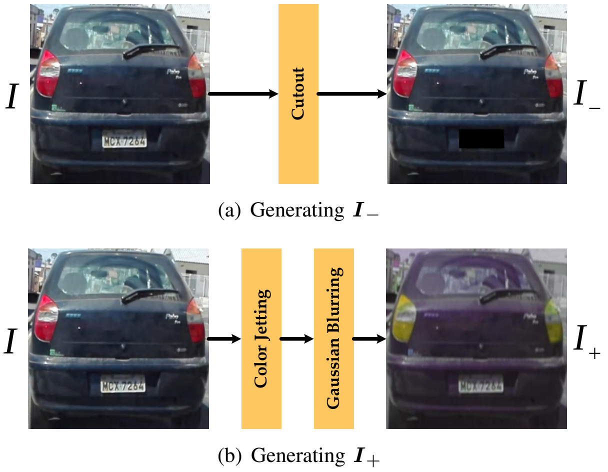
In addition, some researches combine the contrastive learning into supervised learning and propose the joint training frameworks. And the utilization of ground truth information is more popular in many computer vision tasks, especially in pixel-level task like semantic segmentation. Khosla *et al.* [62] propose Supervised Contrastive Learning (SCL) which use labeled data to build the contrastive pairs for image classification, and the proposed supervised contrastive loss achieves better performance compared to supervised learning with cross-entropy loss. Meanwhile, Gunel *et al.* [63] introduce the supervised contrastive learning into natural language understanding classification. They consider that the supervised contrastive learning loss pushes samples from the same class close and samples from different classes further apart in training. Chen *et al.* [64] introduce contrastive learning into video object segmentation, and they use the ground truth masks information to build the foreground region and background region for contrast. Wang *et al.* [65]optimize the contrastive loss between ground truth mask and pseudo mask to facilitate the learning of class-agnostic mask segmentation model. In this paper, we also utilize a supervised contrastive learning method guided by ground truths to decouple the backgrounds and foregrounds in LP detection.

## III. Methods

In this section, the proposed CLPD is described. First, the configuration of contrastive triad is introduced in Section III-A. The whole pipeline of proposed CLPD is briefly described in Section III-B. Then, the details about the contrastive learning in CLPD is shown in Section III-C. Section III-D assumes the changes in latent feature space to explain the effectiveness of contrast, and the experiments verify the correctness of this assumption. Section III-E explains other details about the LP detector.

### A. Contrastive Triad

The core idea of contrastive learning is to learn how to maximize the similarity in similar samples and minimize the similarity in dissimilar samples. To achieve the contrast, we should build a contrastive triad which contains original instance, similar instance, and dissimilar instance. Meanwhile, the keypoint in contrast for detection is to decouple the foregrounds and backgrounds. Intuitively, an instance from LP detection dataset is considered as original instance $\boldsymbol{I}$. The contrast in LP detection task aims to distinguish the LPs from complex street scenes. Thus, the negative instance should not have LPs but contain background scene, and the CutOut [66] augmentation is applied on $\boldsymbol{I}$ to remove the LPs, generating dissimilar instance $\boldsymbol{I}_-$ (Fig. 2(a)) which denotes backgrounds. The CutOut regions are decided by annotated ground truths, so the contrastive learning method in this paper is a supervised contrastive learning method [62] which exploits the information about labeled data. Meanwhile, a positive instance is needed in contrastive learning to constrain the feature space and avoid the collapse of training. The positive instance in contrastive learning should be similar to original instance and it is treated as anchor in contrast, so that a degeneration strategy including color jetting and Gaussian Blurring is implemented on $\boldsymbol{I}$ to acquire an analogous weak augmented image $\boldsymbol{I}_+$ (Fig. 2(b)). There is no translation in $\boldsymbol{I}_+$ during transformation to avoid the disturbances in localization. On the whole, we build a contrastive triad $\mathcal{C}$ which includes original instance $\boldsymbol{I}$, positive instance $\boldsymbol{I}_+$, and negative instance $\boldsymbol{I}_-$ for contrastive learning.



(a) Generating $\boldsymbol{I}_-$

(b) Generating $\boldsymbol{I}_+$

Fig. 2. The constitution of contrastive triad.

This triad aims to decouple the foregrounds and backgrounds, boosting the feature expression of LPs and extracting more discriminative features. The separation between $\boldsymbol{I}$ and $\boldsymbol{I}_-$ prefers to learn the specificity of foregrounds and decouple the foregrounds and backgrounds. Meanwhile, the aggregation between $\boldsymbol{I}$ and $\boldsymbol{I}_+$ aggregates the foreground features in latent space and widens the decision boundary between foregrounds and others, learning discriminative features for detection. In addition, to explore the relation between foregrounds and backgrounds and verify the effectiveness of above contrastive triad $\mathcal{C}$, we compared several different generation strategies for $\boldsymbol{I}_-$ in Section V-A to guide the design of contrastive triad.

### B. Pipeline

The pipeline of proposed CLPD is shown in Fig. 3. The detection efficiency is important in traffic scenario, so that we choose a high-performance one-stage detector, FCOS [16], as baseline in this paper. The reason why we choose FCOS [16] is because its architecture is simple and suitable for the analysis of effectiveness to ensure the improvement are gained from the proposed contrastive strategy. Meanwhile, recent LP detectors [26], [31], [34], [35] are usually based on YOLO [10], YOLOv2 [11], and YOLOv3 [12]. These simple and stable detectors are suitable for industry application. In addition, this paper attempts to design a joint training framework of contrastive learning and supervised learning. Thus, we first introduce this framework into a simple detector to verify the effectiveness. And the proposed CLPD mainly focus on the contrastive learning from backbone features and has the ability to introduce to other detectors (Section V-B). And there are two main branches, contrastive learning branch and supervised learning branch.

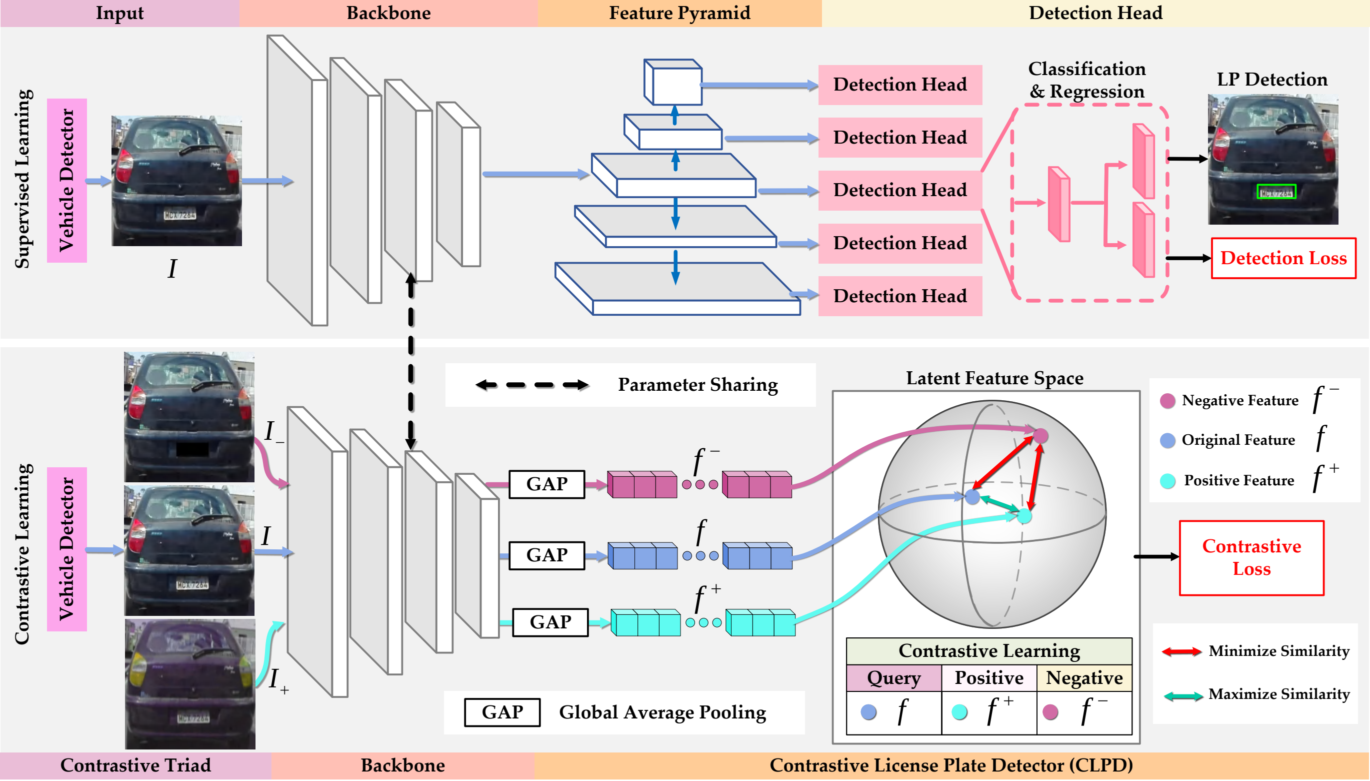


Fig. 3. The pipeline of the proposed CLPD method.

First, a vehicle detector localizes the vehicles in image and crops the vehicle patch $\boldsymbol{I}$. The patch $\boldsymbol{I}$ is used to build the contrastive triad $\mathcal{C}$ according to Section III-A. Then, the contrastive triad $\mathcal{C} = \{\boldsymbol{I}, \boldsymbol{I}_{+}, \boldsymbol{I}_{-}\}$ is input to contrastive learning branch to make the backbone learn the specificity of LP features via contrasting among $\mathcal{C}$. The backbone encodes the $\boldsymbol{I}$, $\boldsymbol{I}_{+}$, and $\boldsymbol{I}_{-}$ into low-dimensional features. And the contrastive loss is calculated with these three features.

As for supervised learning branch, it uses a parameter-sharing backbone with contrastive learning branch to make sure the prior knowledge learned from contrastive learning can be remembered and the parameter-sharing backbone also reduces the computation complexity to achieve high efficiency. And only the original instance $\boldsymbol{I}$ is fed into this branch for the training of detection. Then, the features are input to downstream Feature Pyramid Network (FPN) [67] and a shared detection head among different feature scales to calculate detection loss.

In training, the contrastive learning branch and supervised learning branch are trained jointly. The contrastive learning branch provides more discriminative features, and supervised learning branch achieves the LP detetcion task. In inference, only the supervised learning branch works for LP detetcion, and the contrastive learning branch will not increase any auxiliary computation burden.

### C. Contrastive Learning Branch

The backbone $\mathcal{F}$ processes items in contrastive triad $\mathcal{C} = \{\boldsymbol{I}, \boldsymbol{I}_{+}, \boldsymbol{I}_{-}\}$ and gains the corresponding feature maps. To reduce the computational complexity, a global average pooling layer squeezes the feature maps from $i$-th layer into vectors $\boldsymbol{f}_i$, $\boldsymbol{f}_i^{+}$, and $\boldsymbol{f}_i^{-}$, illustrating in Eq. 1:

$$\begin{aligned} \boldsymbol{f}_i &= GAP(\mathcal{F}(\boldsymbol{I})), \\ \boldsymbol{f}_i^{+} &= GAP(\mathcal{F}(\boldsymbol{I}_{+})), \\ \boldsymbol{f}_i^{-} &= GAP(\mathcal{F}(\boldsymbol{I}_{-})), \end{aligned} \tag{1}$$

where $GAP$ is the global average pooling.

The contrastive learning strategy in proposed CLPD is based on features $\boldsymbol{f}_i$, $\boldsymbol{f}_i^{+}$, and $\boldsymbol{f}_i^{-}$. Many contrastive learning methods [5], [52], [54] choose InfoNCE loss [5] as their contrastive loss, which proves the potential of this loss. In this paper, we also select InfoNCE loss [5] as contrastive loss for the contrast among $\boldsymbol{f}_i$, $\boldsymbol{f}_i^{+}$, and $\boldsymbol{f}_i^{-}$, defining in Eq. 2:

$$\mathcal{L}_q(q, k_{+}, k_{-}) = -\log \frac{exp(q \cdot k_{+}/\tau)}{\Sigma_{i=0}^{K} exp(q \cdot k_i/\tau)}, \tag{2}$$

where $\mathcal{L}_q$ is the InfoNCE loss [5]. $\tau$ is a temperature, a hyper-parameter in $\mathcal{L}_q$. The $q$ is the query feature. There are a single positive key feature $k_{+}$ (*i.e.* $k_0$) and $K$ negative key features $k_{-}$ (*i.e.* $k_i$, $(i = 1, \ldots, K)$). The optimization of $\mathcal{L}_q$ matches the query $q$ with the positive key $k_{+}$ and spreads the negative keys $k_{-}$ out, achieving the aggregation of similar samples and the separation of dissimilar samples.

The purpose of CLPD is to distinguish the foregrounds (*i.e.* LPs) from backgrounds. Therefore, the query $q$ and positive key $k_{+}$ should contain the LPs, and the negative keys $k_{-}$ for backgrounds should have no LP. $\boldsymbol{f}_i$ and $\boldsymbol{f}_i^{+}$ have the representations of LP, so that we choose original feature $\boldsymbol{f}_i$ as query feature and $\boldsymbol{f}_i^+$ as positive feature. The negative keys $k_-$ are $\boldsymbol{f}_i^-$ in the mini-batch, and $K$ equals to the batch size. Therefore, the contrastive learning is defined as Eq. 3:

$$\mathcal{L}_i^{nce} = \mathcal{L}_q(\boldsymbol{f}_i, \boldsymbol{f}_i^+, \boldsymbol{f}_i^-). \tag{3}$$

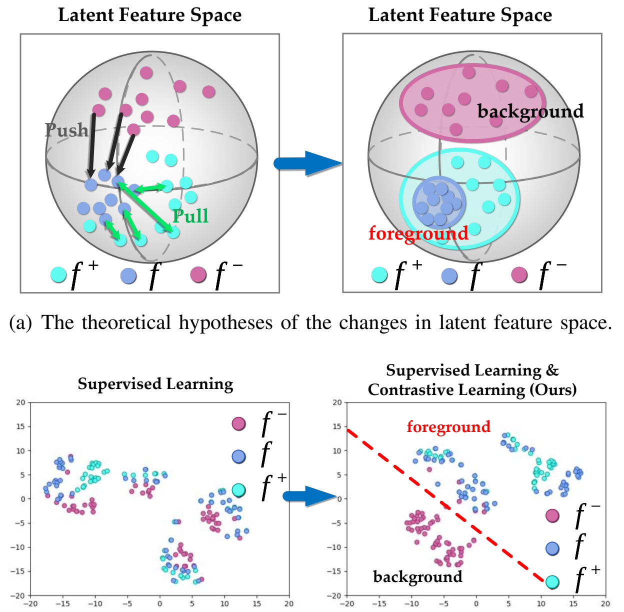
The optimization of contrastive learning makes detector match $\boldsymbol{f}_i$ with $\boldsymbol{f}_i^+$. The augmentation in $\boldsymbol{f}_i^+$ increase the diversity of features, so that the matching of $\boldsymbol{f}_i$ with $\boldsymbol{f}_i^+$ facilitates the detector to learn diverse LP features, increasing the robustness and generalization of detector. Meanwhile, the contrast also spreads $\boldsymbol{f}_i^-$ away from $\boldsymbol{f}_i$, increasing the specificity of LP-related features and boosting the feature expression ability of backbone.

In addition, many detectors introduce the FPN [67] to fuse multi-scale features in detection to adapt the small or large object detetcion. Thus, the contrastive learning branch needs to be exerted on multi-scale features. The total contrastive loss $L_{ctr}$ for multi-scale features is defined in Eq. 4:

$$\mathcal{L}_{ctr} = \Sigma_{i=1}^{L} \lambda_i \mathcal{L}_i^{nce}, \tag{4}$$

where $L$ is the total number of layers used in contrast. The $\lambda_i$ is the weighting parameters for loss of features from $i$-th layer, and $\mathcal{L}_i^{nce}$ is the loss (*i.e.* Eq. 3) for $i$-th layer features. To take ResNet [68] as an example, $L = 3$ and the multi-scale features from the 2-nd, 3-rd, and 4-th layers (denoted as $Conv3\_x$, $Conv4\_x$, and $Conv5\_x$ in [68]) are used to calculate $\mathcal{L}_{ctr}$ for contrastive learning.

### D. Effect Analysis



(a) The theoretical hypotheses of the changes in latent feature space.

(b) The experimental verification of the decoupling of backgrounds and foregrounds.

Fig. 4. The foreground feature aggregation processing in CLPD.

The foreground and background features will be decoupled via the contrastive learning in CLPD. We suppose the changes in latent feature space as Fig. 4(a). Without the training of contrastive learning, the features of $\boldsymbol{f}, \boldsymbol{f}^+, \boldsymbol{f}^-$ are distributed in latent feature space chaotically, and there is no obvious boundary between foreground features (*i.e.* $\boldsymbol{f}$ and $\boldsymbol{f}^+$) and background features (*i.e.* $\boldsymbol{f}^-$), leading to the misunderstanding of object information. When the contrastive learning branch is introduced into joint training, the similarity between $\boldsymbol{f}$ and $\boldsymbol{f}^+$ is maximized and the similarity between $\boldsymbol{f}$ and $\boldsymbol{f}^-$ is minimized. This is means that the background features $\boldsymbol{f}^-$ will push the foreground features $\boldsymbol{f}$ away, and the features of positive sample similar to foreground $\boldsymbol{f}^+$ will pull the foreground feature $\boldsymbol{f}$ together, which decouples the foregrounds and backgrounds and aggregates the foreground feature $\boldsymbol{f}$.

To verify whether the changes in latent feature space is same as the above assumption, we first extract the local feature maps of annotated LP regions in $\boldsymbol{f}, \boldsymbol{f}^+, \boldsymbol{f}^-$ according to ground truth labels. And then the tSNE method [69]is used to visualize the feature distribution in latent space. We select several data from the testing set in UFPR-ALPR dataset [26] for visualization experiment. The visualization results are shown in Fig. 4(b) , every point in figure represents the feature of a LP instance after dimension reduction. We find that the supervised learning method can not separate background features and foreground features clearly. The $\boldsymbol{f}$ and $\boldsymbol{f}^-$ are mixed in latent feature space in the left figure of Fig. 4(b). When we introduce the contrastive learning into LP detection, the proposed CLPD distinguish the background feature $\boldsymbol{f}^-$ and foreground feature $\boldsymbol{f}$ effectively via contrast among contrastive triad $\mathcal{C}$ (the right figure of Fig. 4(b)), and there is a distinct decision boundary between backgrounds and foregrounds. This evidence proves that the proposed CLPD method decouples the foreground and background effectively, aggregating the foreground features and increasing the expression of extracted foreground features.

### E. Other Details

The contrastive learning brach is introduced above, but the main task in this paper is LP detection. Therefore, the supervised learning branch for detection training is necessary. The supervised learning branch is traditional detection pipeline based on one-stage detector. The input vehicle patch $\boldsymbol{I}$ is first sent to backbone to extract features. And then the multi-scale features are fused by FPN [67] to enhance the capability of representations. The output features from FPN [67] are processed by a shared detection head for category classification and bounding box regression. The training of one-stage detector mainly depends on the optimization of detection loss $\mathcal{L}_{det}$ under supervised learning. $\mathcal{L}_{det}$ has three core parts, classification loss $\mathcal{L}_{cls}$, regression loss $\mathcal{L}_{reg}$, and auxiliary loss $\mathcal{L}_{other}$. The supervised learning branch and contrastive learning branch are trained jointly. Thus, the total loss $\mathcal{L}_{total}$ is defined as follow:

$$\begin{aligned} &\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{other}, \\ &\mathcal{L}_{total} = \mathcal{L}_{det} + \eta \mathcal{L}_{ctr}, \end{aligned} \tag{5}$$

where $\mathcal{L}_{cls}$ is the cross-entropy loss or focal loss [18] for classification, and $\mathcal{L}_{reg}$ is GIoU loss [70] to localize the bounding box. There are some auxiliary loss $\mathcal{L}_{other}$ to further improve the detection performance, such as centerness loss in FCOS [16]. The $\mathcal{L}_{ctr}$ is the loss for contrastive learning (*i.e.* Eq. 4), and it has a hyper-parameter $\eta$ to balance the terms in total loss $\mathcal{L}_{total}$.

## IV. Experiments

This section shows the experimental results of the proposed CLPD method. Section IV-A first introduces the evaluation metrics in detection. The datasets used in experiments are introduced in Section IV-B. The experimental details are described in Section IV-C. Section IV-D shows the experimental results on several popular LP detection datasets. Finally, Section IV-E discusses the selection of hyper-parameters in our proposed CLPD.

### A. Evaluation Metrics

As for detection task, there are several general quantitative criteria for performance evaluation. The predictions are compared with the ground truths. In comparison, the True Positive (TP), False Positive (FP), and False Negative (FN) are counted to calculate the Precision (P) and Recall (R). Meanwhile, the Intersection over Union (IoU) is calculated to classify the predictions into TP, FP, and FN. To comprehensively consider the precision and recall, F1-score and Average Precision (AP) are proposed. The F1-score calculated by P and R, and the AP is the area under P-R curve. The above general evaluation metrics for detection task are illustrated in Eq. 6:

$$\begin{aligned}
&P=\frac{TP}{TP+FP},\\
&R=\frac{TP}{TP+FN},\\
&IoU(\boldsymbol{pred},\boldsymbol{gt})=\frac{\boldsymbol{pred}\cap\boldsymbol{gt}}{\boldsymbol{pred}\cup\boldsymbol{gt}},\\
&F1\text{-}score=\frac{2\times R\times P}{P+R},\\
&AP=\int_0^1 p(r)dr,
\end{aligned}\tag{6}$$

where $\boldsymbol{pred}$ is the predicted bounding box and $\boldsymbol{gt}$ is the ground truth bounding box. The $p(r)$ is the P-R curve. The F1-score and AP are the most important metrics in detection task, and higher F1-score and AP denote better performance.

### B. Datasets

In this paper, the proposed CLPD is evaluated on three LP detection datasets, UFPR-ALPR, CCPD, SSIG-SegPlate.

**UFPR-ALPR**: This dataset is a video LP detection dataset for Brazilian LPs. It contains 150 video clips (4,500 frames) captured on moving vehicles. The video clips have a fixed length of 30 frames. All frames have a fixed resolution of $1,920 \times 1,080$ pixels. The training set and testing set both include 60 videos, and the validation set contains 30 videos. Every video clip only contains a single annotated LP. There are three types LPs in UFPR-ALPR, *i.e.* gray LP, red LP, and motorcycle LP.

**CCPD**: This dataset is a Chinese LP detection dataset. It contains $250,000$ images captured in parking lots. All images have a fixed resolution of $720 \times 1160$. CCPD has nine subset for different scene: CCPD-Base ($200,000$ images), CCPD-DB ($20,000$ images), CCPD-FN ($20,000$ images), CCPD-Blur (more than $20,000$ images), CCPD-Rotate ($10,000$ images), CCPD-Tilt ($10,000$ images), CCPD-Weather ($10,000$ images), CCPD-Challenge ($10,000$ images), and CCPD-NP (more than $3,000$ images). In experiments, we only choose CCPD-Base as dataset to train the detector and test performance. $100,000$ images in CCPD-Base are used for training, and the rest $100,000$ images are used for testing.

**SSIG-SegPlate**: This dataset is a Brazilian LP detection dataset which has $2,000$ images captured by a fixed static camera. All images have a size of $1,920 \times 1,080$ pixels and contain a single annotated LP. There are passenger vehicle LP, bus or truck LP, and motorcycle LP in this dataset. The training set and testing set both include 800 images, and the validation set has 400 images.

### C. Implementation Details

The proposed CLPD is a common framework for one-stage detector. Thus, we select a prevalent high-performance one-stage detector, FCOS [16], as the baseline to verify the effectiveness of the proposed CLPD. The code is available at https://github.com/Dinghaoxuan/CLPD.

**Inputs**: The input image is directly resized from $1,920 \times 1,080$ to $512 \times 512$. And this resized image is fed into a pre-trained vehicle detector to detect vehicles in image first. This vehicle detector is provided by [26] and has 100.0 Recall and 125 FPS on UFPR-ALPR dataset [26] to ensure all vehicles in images are detected. The Precision of vehicle detector is 93.7 but it is unimportant because the vehicle detector should focus on Recall rather than Precision to make sure that all possible patches with LPs can be fed to subsequent LP detector. In training, vehicle patches are utilized to generate contrastive triad (Section III-A) for contrastive learning branch. And the contrastive triad is input to detector for joint training of supervised learning and contrastive learning. In inference, the vehicle patches are directly input to detector for LP detection, because there is no contrastive learning branch in inference.

**Training Strategy**: The backbone is a pre-trained ResNet-18 on ImageNet from model zoo of Pytorch [81]. We jointly train the contrastive learning branch and supervised learning branch in total 30 epochs. We use Stochastic Gradient Descent (SGD) method as optimizer to train the detector. The warm-up learning rate is $10^{-4}$ for the first $5$ epochs. After warm-up training, the learning rate is set to $10^{-3}$. After 10 training epochs, the learning rate decreases to $10^{-4}$. And after 20 epochs, the learning rate further reduce to $10^{-5}$.

**Environment:** We use an equipment with Intel(R) CPU Core(TM) i7-6900K @ 3.4GHz, 64GB RAM, and a NVIDIA GTX1080TI GPU for all experiments. And the models are built by Pytorch framework [81].

### D. Performance

Fig 5 show some detection results on three LP detection datasets. Table I illustrates the quantitative results on three

TABLE I
THE COMPARISON RESULTS OF THE PROPOSED CLPD. THE RESULTS ARE COMPREHENSIVE PERFORMANCE OF CLPD WHICH CONTAINS VEHICLE DETECTION AND LP DETECTION. THE PROPOSED CLPD ACHIEVES BETTER PERFORMANCE COMPARED TO LISTED LP DETECTION METHODS.

| Method | Dataset | Backbone | P | R | F1-score | AP | FPS |
|---|---|---|---|---|---|---|---|
| FCOS (Baseline) [16] | UFPR-ALPR | ResNet-18 | 93.6 | 96.7 | 95.1 | 92.9 | 30.8 |
| CLPD (Ours) | | ResNet-18 | 97.2 | 100.0 | **98.6** (+3.5) | **97.9** (+5.0) | 30.8 |
| YOLOv2 [11] | UFPR-ALPR | Darknet-19 | 92.7 | 94.7 | 93.7 | 87.8 | 58.9 |
| YOLOv3 [12] | | Darknet-53 | 94.9 | 97.4 | 96.1 | - | 47.6 |
| YOLOv4 [13] | | CSPDarknet-53 | 93.1 | 92.6 | 92.8 | 88.5 | 19.8 |
| EAST [71] | | PVANet | 92.8 | 99.9 | 96.2 | - | 38.5 |
| R-FCN [72] | | ResNet-101 | 94.6 | 99.8 | 94.5 | - | 15.1 |
| Laroca *et al.* [26] | | FAST-YOLO | - | 98.3 | - | - | 66.5 |
| FGFA [73] | | ResNet-50 | 97.2 | 98.3 | 97.7 | - | 11.5 |
| Lee *et al.* [38] | | ResNet-50 | - | 99.2 | - | - | 14.0 |
| FCOS (Baseline) [16] | CCPD-Base | ResNet-18 | 99.3 | 100.0 | 99.6 | 99.4 | 35.6 |
| CLPD (Ours) | | ResNet-18 | 99.8 | 100.0 | **99.9** (+0.3) | **99.8** (+0.4) | 35.6 |
| Cascade Classifier [74] | CCPD-Base | - | 55.4 | - | - | 47.2 | 32.0 |
| SSD300 [45] | | VGG-16 | 99.1 | - | - | 94.4 | 40.0 |
| YOLOv2 [11] | | Darknet-19 | 98.8 | - | - | 93.1 | 42.0 |
| YOLOv4-Tiny [13] | | CSPDarknet-53-Tiny | 94.7 | - | - | 90.2 | 110.2 |
| YOLOv4 [13] | | CSPDarknet-53 | 99.1 | - | - | 98.8 | 33.1 |
| Faster R-CNN [8] | | VGG-16 | 98.1 | - | - | 92.9 | 15.0 |
| TE2E [75] | | VGG-16 | 98.5 | - | - | 94.2 | 3.0 |
| RPnet [76] | | - | 99.3 | - | - | 94.5 | 61.0 |
| SLPNet [77] | | ShuffleNetv2 | 99.9 | 99.3 | 99.6 | - | 25.0 |
| Silva and Jung [40] | | Darknet-19 | 86.1 | - | - | - | 13.6 |
| Pham [78] | | - | 99.3 | - | - | 98.1 | 168.3 |
| Lee *et al.* [38] | | ResNet-50 | 96.1 | - | - | - | 14.0 |
| FCOS (Baseline) [16] | SSIG-SegPlate | ResNet-18 | 96.4 | 99.9 | 98.1 | 97.7 | 37.5 |
| CLPD (Ours) | | ResNet-18 | 99.7 | 98.5 | **99.1** (+1.0) | **98.4**(+0.7) | 37.5 |
| Silva and Jung [79] | SSIG-SegPlate | FAST-YOLO | 95.1 | 99.5 | 97.3 | - | 8.7 |
| Laroca *et al.* [26] | | FAST-YOLO | - | 100.0 | - | - | - |
| Silva and Jung [80] | | FAST-YOLO | 95.1 | 99.5 | 97.3 | - | 8.7 |
| Laroca *et al.* [32] | | Fast-YOLOv2 | 95.3 | 99.8 | 97.5 | - | 29.4 |

datasets in Section IV-B. All of the evaluation metrics of proposed CLPD are comprehensive performance on two steps (*i.e.* vehicle detection and LP detection in Section III-B). The undisputed metrics in LP detection is F1-score and AP introduced in Section IV-C. Thus, we mainly focus on the comparison on F1-score and AP to evaluate the performance.

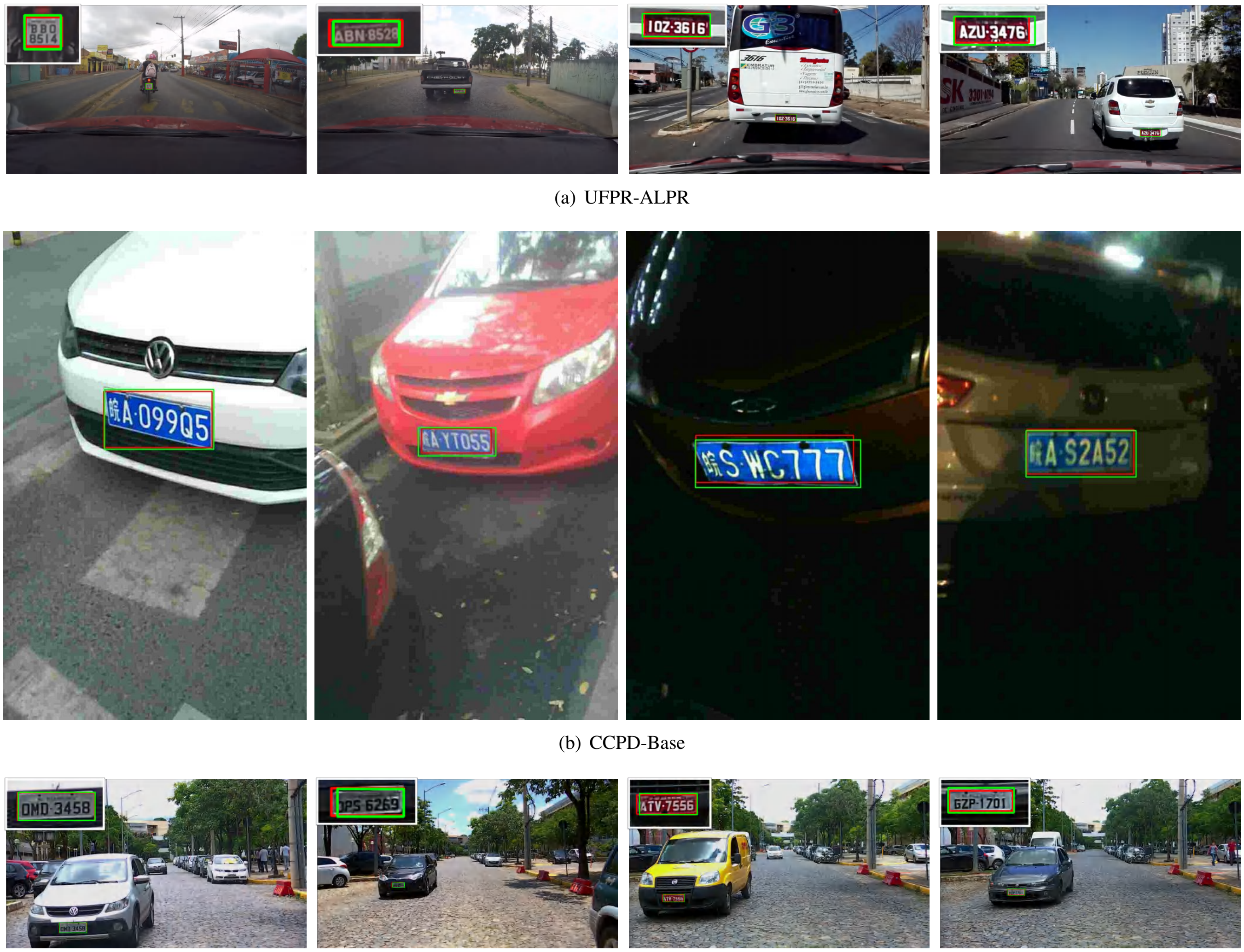
**UFPR-ALPR**: Compared to the baseline (FCOS [16]), our proposed CLPD achieves significantly improvement, there is an increase of 3.5 on F1-score and an increase of 5.0 on AP. Compared to other LPD methods, the proposed CLPD also reaches better performance (98.6 on F1-score and 97.9 on AP). When we compare the results between CLPD and FGFA [73], we find that the CLPD with ResNet-18 backbone even exceeds the FGFA with ResNet-50 backbone. Besides, the proposed CLPD reaches 100.0 on Recall which exceeds the contrast LP detector proposed by Lee *et al.* [38] (99.2 on Recall). This means the proposed contrastive learning branch in CLPD boosts the feature expression ability of light-weight backbone successfully and extracting more discriminative features, improving the LP detection performance.

**CCPD-Base**: On CCPD-Base dateset, the experiment results also show that the proposed CLPD prompts the detection accuracy of the baseline. From Table I, we find that the LPD methods have already reached promising performance on CCPD-Base, but some hard samples in dataset still restrict the imporvement of accuracy. The reason why there are blank in Table I is because that the general evaluation metric for CCPD dataset is Accuracy (*i.e.* Precision). The most of comparison methods only provide Precision rather than Recall and F1-score in their papers. Thus, we only compare the Precision on this dataset. Our proposed CCPD tackles these hard samples effectively and achieves highest value on F1-score (99.9) and AP (99.8). Meanwhile, the CLPD also has competitive result on Precision (P), which is only 0.1 smaller than SLPNet [77]. Although the Precision (P) of CLPD is slightly lower than that of SLPNet [77], the F1-score of CLPD is higher than SLPNet [77] notably. In addition, compared with the Lee *et al.* [38], our proposed CLPD has higher accuracy on CCPD-Base dataset. Comprehensively, our proposed CLPD achieves promising performance on CCPD-Base dataset.

**SSIG-SegPlate**: The proposed CLPD obtains an increase of 1.0 on F1-score and 0.7 on AP compared with the baseline. Meanwhile, the performance of CLPD also exceeds other methods in comparison distinctly. This also proves the effectiveness of the proposed CLPD.

## *E. Ablation Studies*

To explore how to promote the performance, some ablation studies on UFPR-ALPR dataset [26] are implemented to further explore a better parameter configuration for CLPD.



(a) UFPR-ALPR

(b) CCPD-Base

(c) SSIG-SegPlate

Fig. 5. The detection results of CLPD on three datasets. The **red boxes** are ground truths, and **green boxes** are predicted bounding boxes.

TABLE II
THE COMPARISON RESULTS OF DIFFERENT $\lambda_i$ IN EQ. 4 ON UFPR-ALPR DATASET [26].

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | P | R | F1-score | AP |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 97.0 | 99.8 | 98.4 | 97.4 |
| 0 | 1 | 0 | 92.6 | 95.6 | 94.1 | 89.3 |
| 0 | 0 | 1 | 96.6 | 98.9 | 97.8 | 97.1 |
| 1 | 1 | 0 | 95.9 | 98.8 | 97.3 | 94.9 |
| 1 | 0 | 1 | 98.0 | 98.8 | 98.4 | 96.9 |
| 0 | 1 | 1 | 93.7 | 96.9 | 95.3 | 90.8 |
| 1 | 1 | 1 | 97.0 | 99.3 | 98.1 | 97.6 |
| 2 | 1 | 1 | 95.3 | 98.8 | 97.1 | 95.2 |
| 1 | 2 | 1 | 93.2 | 98.4 | 95.7 | 92.2 |
| 1 | 1 | 2 | 97.2 | 100.0 | **98.6** | **97.9** |

**Multi-scale Feature Weighting:** The multi-scale features are frequently used in recent detectors. The proposed CLPD also contrasts the multi-scale features to fully exploit contrastive learning. In Eq. 4, the InfoNCE loss of $i$-th layer $\mathcal{L}_i^{nce}$ are weighted by a hyper-parameter $\lambda_i$. To explore the importance of features with different scales, Table II demonstrates the comparison results.

In the comparison among $\lambda_i = \{1, 0, 0\}$, $\lambda_i = \{0, 1, 0\}$, and $\lambda_i = \{0, 0, 1\}$, we find that the configurations of $\lambda_i = \{1, 0, 0\}$ and $\lambda_i = \{0, 0, 1\}$ reach higher performance. This means the low-level feature from 2-nd layer ($Conv3\_x$) and high-level feature from 4-th layer ($Conv5\_x$) have more influence on LP detection task. The LPs are small in image and the low-level feature has high resolution which is suitable for small object detection. The high-level feature has a large receptive field to correlate much information from image content for accurate localization. The fusion of features from 2-nd layer, 3-rd layer, and 4-th layer ( *i.e.* $Conv3\_x$, $Conv4\_x$, and $Conv5\_x$) (*i.e.* $\lambda_i = \{1, 1, 1\}$) achieves better performance compared to the configurations of $\lambda_i = \{1, 1, 0\}$, $\lambda_i = \{1, 0, 1\}$, and $\lambda_i = \{1, 0, 1\}$. This proves that the fusion of conrtastive learning on multi-scale features are beneficial to detection. Meanwhile, we find that the contrastive learning on different scales has different contributions, so that we further emphasize the weightings of features from different layers. Eventually, the configuration of $\lambda_i = \{1, 1, 2\}$ achieves the best performance.

**Comparison of different** $\eta$: To balance the losses in $\mathcal{L}_{total}$

TABLE III
THE COMPARISON RESULTS OF DIFFERENT $\eta$ IN EQ. 5 ON UFPR-ALPR DATASET [26].

| $\eta$ | P | R | F1-score | AP |
|---|---|---|---|---|
| 0.01 | 97.0 | 97.6 | 97.3 | 94.7 |
| 0.10 | 97.2 | 100.0 | **98.6** | **97.9** |
| 1.00 | 93.8 | 93.0 | 93.4 | 87.3 |

(Eq. 5), we introduce a hyper-parameter $\eta$ to adjust magnitude of contrastive loss $\mathcal{L}_{ctr}$ (Eq. 4). Table III shows the comparison results about $\eta$. When the $\eta$ is 0.01, $\mathcal{L}_{ctr}$ is too small and has limited contribution to improvement. On the contrary, when the $\eta$ is 1, the optimization of $\mathcal{L}_{total}$ is inclined to minimize the $\mathcal{L}_{ctr}$ rather than $\mathcal{L}_{det}$, causing the misleading of optimization and decreasing the detection performance.

## V. DISCUSSION

In this section, we discuss some issues to explain the reason why CLPD achieves better performance. The experiments are all implemented on the most challenging dataset, UFPR-ALPR dataset [26], to show the effectiveness credibly.

### A. Design of contrastive triad

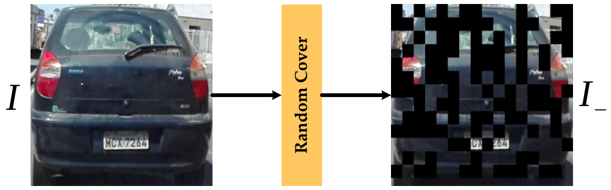


Fig. 6. The random cover strategy to generate $\boldsymbol{I}_{-}$.

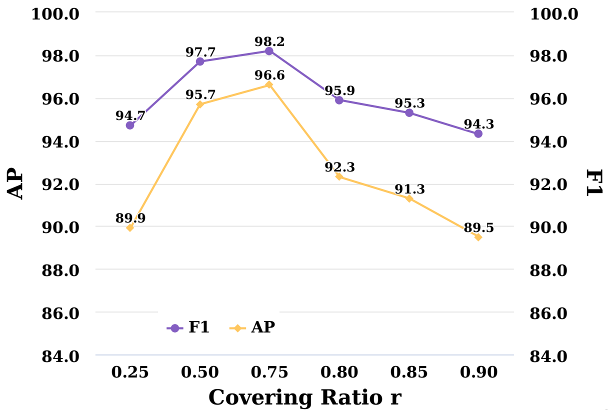


Fig. 7. The influence of random covering ratio $r$.

To explore how to efficiently decouple the foregrounds and backgrounds, we attempt to find out the influence of foregrounds and backgrounds. Thus, we design a random cover strategy to randomly remove the content patch in image with a covering ratio $r$ which is shown in Fig. 6. The vehicle patch is split to $16 \times 16$ small patches, and then a percentage of small patches are randomly selected and removed from the image according to the covering ratio $r$. We adjust the covering ratio $r$ and analyze the influence of removed contents. Fig. 7 demonstrates the detection performance with different $r$.

From Fig. 7, we find that when the $r$ is small ($r = 0.25$) the contrastive learning can not boost the performance. This is because that the random covering do not remove the foregrounds (*i.e.* LPs) from negative sample and the LPs information in negative sample confuses the learning of backbone. With the increasing of $r$, the detection performance increases until $r = 0.75$. However, the increasing of $r$ after $r = 0.75$ does harm to the detection accuracy. This is because a high covering ratio $r$ removes the most of content which restricts backbone to learning the backgrounds information. The reason why the detection performance is still over 90.0 AP when $r > 0.8$ is because the contrast is meaningless due to lack of content, and the contrast starts to inhibit the improvement of detection accuracy, but the supervised learning remains effective which maintains a high level of detection accuracy (95.1 F1-score and 92.9 AP for supervised learning baseline). The highest performance in Fig. 7 is still lower than the proposed CLPD with triad in Section III-A. From this phenomenon, we find that the negative sample should eliminate the foreground information and preserve the background information as much as possible to build effective contrast. Therefore, we decide use the CutOut method and ground truth information to remove the LPs in images.

### B. Comparison with different baselines

To further explore the generalization of the proposed CLPD methods, we select different baselines and introduce the contrastive learning branch in Section III-C into them to test the improvements.We choose several prevalent one-stage detectors, FCOS [16], YOLOv2 [11], and YOLOv4 [13], in this comparison.

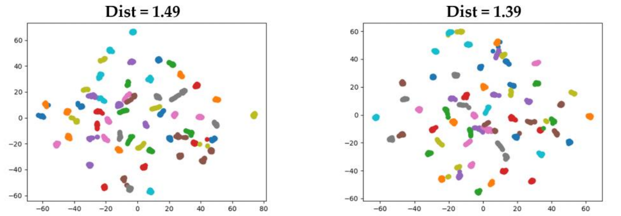
TABLE IV
THE COMPARISON RESULTS OF DIFFERENT BASELINES ON UFPR-ALPR DATASET [26].

| Method | P | R | F1-score | AP |
|---|---|---|---|---|
| FCOS [16] (Baseline) | 93.6 | 96.7 | 95.1 | 92.9 |
| FCOS-CLPD (Ours) | 97.2 | 100.0 | **98.6** (+3.5) | **97.9** (+5.0) |
| YOLOv2 [11] (Baseline) | 92.7 | 94.7 | 93.7 | 87.8 |
| YOLOv2-CLPD (Ours) | 98.4 | 93.5 | **95.9** (+1.2) | **91.5** (+3.7) |
| YOLOv4 [13] (Baseline) | 93.1 | 92.6 | 92.8 | 88.5 |
| YOLOv4-CLPD (Ours) | 97.4 | 94.4 | **96.1** (+3.3) | **92.6** (+4.1) |

The details of training configurations are as follows.All of the inputs for detector are the resized images with $512 \times 512$ pixels, and the batchsize is 4 for all detectors which is described in Section IV-C. As for YOLOv2, the backbone is Darknet-19 [11], and there is no pre-training for YOLOv2. The YOLOv2 is optimized by SGD. The initial parameter of YOLOv2 is random sampled from a Gaussian distribution $\mathcal{N}(0, 0.01)$. The YOLOv2 is totally trained by 160 epochs. The first 5 epochs of YOLOv2 have a learning rate of $10^{-4}$, and after 5 epochs, the learning rate increases to $5 \times 10^{-4}$ during 75 epochs. Then, the subsequent 30 epochs have a learning rate of $10^{-4}$. Finally, the last 50 training epochs have the learning rate of $10^{-5}$. As for YOLOv4, the backbone network is CSPDarknet-53 [13], which is directly initialized by random sampling from Gaussian distribution $\mathcal{N}(0, 0.02)$. The YOLOv4 is optimized by Adam optimizer. There are 80 training epochs for YOLOv4. The first 30 training epochs have a learning rate of $10^{-3}$ and the rest 50 training epochs have a learning rate of $10^{-4}$.

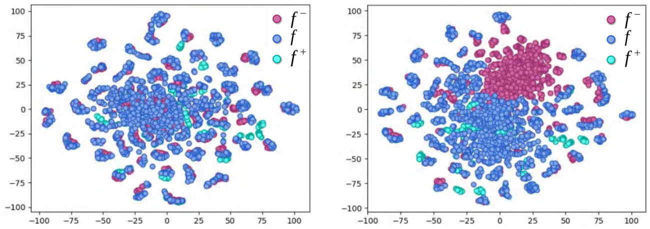
Table IV illustrates the comparison results on UFPR-ALPR dataset [26]. We choose several prevalent one-stage detectors, FCOS [16], YOLOv2 [11], and YOLOv4 [13], in this comparison. From Table IV, we find that the proposed contrastive learning branch in CLPD improves the detection performance significantly on different baselines. This means the proposed contrastive learning strategy is suitable for different LP detectors and has great generalization ability.

### *C. Decoupling of foreground and background*



(a) FCOS(Baseline) (b) CLPD(Ours)

Fig. 8. The visualizations of LP features in UFPR-ALPR dataset [26] through tSNE [69]. The Dist on top is the average intra-cluster distance.



(a) FCOS(Baseline) (b) CLPD(Ours)

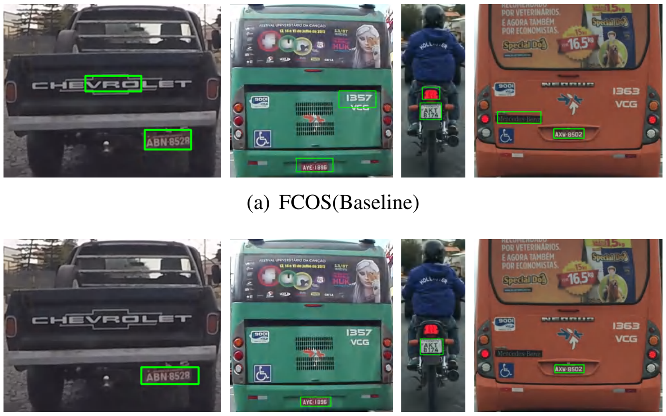
Fig. 9. The visualizations of features in contrastive triad on UFPR-ALPR dataset [26] through tSNE [69].

In Section III-D, we assume that the proposed CLPD method decouples the foregrounds from backgrounds and aggregates LP features in latent space via contrastive learning. To verify this hypothesis and find the reason why CLPD improves detection performance, the visualization of LP features from the 2-nd layer of backbone (*i.e.* $Conv3\_x$) are demonstrated in Fig. 8. Specifically, the vehicle patches from $\boldsymbol{I}$ are first input to the LP detector, and then the feature patches of annotated LP regions are extracted and resized to a fixed size of $256 \times 1 \times 1$ for visualization by tSNE [69]. A cluster in Fig. 8 denotes a video clip (30 frames) in UFPR-ALPR dataset [26]. Because every video clip in UFPR-ALPR dataset only contains a single LP instance, the LP features from the same video clip constitute a cluster in Fig. 8. Therefore, a dense cluster means that the features of a LP instance are aggregated more tightly, and the decision boundary of this LP instance is wide in detection, gaining more discriminative features. The aggregation effectiveness of LPs (*i.e.* foreground) is quantitative evaluated by the average Euclidean distance $\mathrm{Dist}$ which is calculated by Eq. 7:

$$\mathrm{Dist} = \frac{1}{num}\Sigma_{\mathrm{i}=0}^{num}\frac{1}{|\mathbf{c}_\mathrm{i}|}\Sigma_{\mathrm{j}=0}^{|\mathbf{c}_\mathrm{i}|}||\mathbf{p}_\mathrm{ij} - \mathbf{ctr}_\mathrm{i}||^2, \qquad (7)$$

where $\mathbf{C} = \{\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{num}\}$ is the set of clusters after tSNE [69], and $num$ is the number of clusters in $\mathbf{C}$. The point $\mathbf{p}_{ij}$ is the $j$-th points in cluster $\mathbf{c}_i$, and $\mathbf{ctr}_i$ is the cluster center of $\mathbf{c}_i$. A smaller $\mathrm{Dist}$ represents that the features of a single LP instance are tighter in latent feature space. We find that the proposed CLPD achieves denser clusters ($\mathrm{Dist} = 1.39$) in latent space compared to the baseline model ($\mathrm{Dist} = 1.49$), extracting more discriminative features.

The difference of $\mathrm{Dist}$ between CLPD and baseline is slight. In order to further explore the effectiveness of CLPD, we visualize the features in contrast, *i.e.* $\boldsymbol{f}, \boldsymbol{f}^+, \boldsymbol{f}^-$ in Fig. 9. The LP regions are gained from ground truths, and the corresponding feature patches in 2-nd layer features of of backbone from $\boldsymbol{f}, \boldsymbol{f}^+, \boldsymbol{f}^-$ are extracted and resized to a fixed size of $256 \times 16 \times 16$ for visualization. It is obvious that the proposed CLPD separates the background $\boldsymbol{f}^-$ from $\boldsymbol{f}$. The baseline model can not distinguish the $\boldsymbol{f}^-$ from $\boldsymbol{f}$, so that $\boldsymbol{f}^-$ are mixed with $\boldsymbol{f}$. This means the baseline model may misunderstanding the backgrounds as foregrounds, causing false predictions. Therefore, the proposed CLPD effectively alleviates the misjudgments in detection. To prove this opinion, we compare some hard sample with scene text disturbance on baseline and CLPD, which is shown in Fig. 10. Evidently, the baseline model misjudges the scene texts on vehicle body as LPs, but the proposed CLPD successfully filters those misleading texts and alleviates the disturbances from similar scene texts, improving the detection accuracy.



(a) FCOS(Baseline)

(b) CLPD(Ours)

Fig. 10. The detection performance comparison on hard samples in UFPR-ALPR dataset [26].

## VI. Conclusion

This paper proposes a LP detector with contrastive learning, named Contrastive License Plate Detector (CLPD). Specifically, we first design a special contrastive triad which aims to decouple the foregrounds and backgrounds. Based on this triad, a contrastive learning branch and corresponding contrastive loss are introduced into LP detector to promote the feature expression ability of backbone, extracting more discriminative features. The contrastive learning branch for contrast and supervised learning branch for detection are jointly trained in training. In inference, only the supervised learning branch for detection is implemented, so that the contrastive learning branch does not bring any burden in inference, keeping the efficiency of one-stage detectors. The experiments show that the proposed CLPD significantly improves the detection performance of the baselines and reaches better performance compared with other LP detectors. The contrastive learning branch achieves improvement under different baselines, proving the great generalization ability of proposed contrastive learning method. And the visualization results also verify that the CLPD extracts more discriminative features compared to the baseline.

In this paper, the contrastive strategy is only used to decouple foregrounds and backgrounds. This means this contrastive learning strategy is only suitable for the single category detection task, such as LP detection. In the future, we will attempt to introduce more diverse contrastive strategies to extend this work into general object detection task.

## References

[1] C. Liu and F. Chang, "Hybrid cascade structure for license plate detection in large visual surveillance scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2122–2135, 2019.

[2] L. Zhang, P. Wang, H. Li, Z. Li, C. Shen, and Y. Zhang, "A robust attentional framework for license plate recognition in the wild," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 6967–6976, 2021.

[3] M. Molina-Moreno, I. González-Díaz, and F. Díaz-de-María, "Efficient scale-adaptive license plate detection system," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2109–2121, 2019.

[4] M. S. Al-Shemarry, Y. Li, and S. A. Abdulla, "An efficient texture descriptor for the detection of license plates from vehicle images in difficult conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 553–564, 2020.

[5] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[6] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA*. IEEE Computer Society, 2014, pp. 580–587.

[7] R. B. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile*. IEEE Computer Society, 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, Quebec, Canada*, 2015, pp. 91–99.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy*. IEEE Computer Society, 2017, pp. 2980–2988.

[10] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA*. IEEE Computer Society, 2016, pp. 779–788.

[11] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA*. IEEE Computer Society, 2017, pp. 6517–6525.

[12] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[13] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020.

[14] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, vol. 11218. Springer, 2018, pp. 765–781.

[15] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *CoRR*, vol. abs/1904.07850, 2019.

[16] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South)*, 2019, pp. 9626–9635.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3431–3440.

[18] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy*. IEEE Computer Society, 2017, pp. 2999–3007.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 12346. Springer, 2020, pp. 213–229.

[21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[22] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 3601–3610.

[23] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional DETR for fast training convergence," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 3631–3640.

[24] W. Wang, J. Zhang, Y. Cao, Y. Shen, and D. Tao, "Towards data-efficient detection transformers," *CoRR*, vol. abs/2203.09507, 2022.

[25] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.

[26] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "A robust real-time automatic license plate recognition based on the YOLO detector," in *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*. IEEE, 2018, pp. 1–10.

[27] G. R. Gonçalves, S. P. G. da Silva, D. Menotti, and W. R. Schwartz, "Benchmark for license plate character segmentation," *J. Electronic Imaging*, vol. 25, no. 5, p. 053034, 2016.

[28] M. A. Rafique, W. Pedrycz, and M. Jeon, "Vehicle license plate detection using region-based convolutional neural networks," *Soft Comput.*, vol. 22, no. 19, pp. 6429–6440, 2018.

[29] M. Dong, D. He, C. Luo, D. Liu, and W. Zeng, "A cnn-based approach for automatic license plate recognition in the wild," in *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017.

[30] H. Li, P. Wang, and C. Shen, "Towards end-to-end car license plates detection and recognition with deep neural networks," *CoRR*, vol. abs/1709.08828, 2017.

[31] G. Hsu, A. Ambikapathi, S. Chung, and C. Su, "Robust license plate detection in the wild," in *14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy*. IEEE Computer Society, 2017, pp. 1–6.

[32] R. Laroca, L. A. Zanlorensi, G. R. Gonçalves, E. Todt, W. R. Schwartz, and D. Menotti, "An efficient and layout-independent automatic license plate recognition system based on the YOLO detector," *CoRR*, vol. abs/1909.01754, 2019.

[33] X. Kong, K. Wang, M. Hou, X. Hao, G. Shen, X. Chen, and F. Xia, "A federated learning-based license plate recognition scheme for 5g-enabled internet of vehicles," *IEEE Trans. Ind. Informatics*, vol. 17, no. 12, pp. 8523–8530, 2021.

[34] Y. Jamtsho, P. Riyamongkol, and R. Waranusast, "Real-time license plate detection for non-helmeted motorcyclist using YOLO," *ICT Express*, vol. 7, no. 1, pp. 104–109, 2021.

[35] M. Shahidi Zandi and R. Rajabi, "Deep learning based framework for iranian license plate detection and recognition," *Multimedia Tools and Applications*, vol. 81, no. 11, pp. 15 841–15 858, 2022.

[36] S. M. Silva and C. R. Jung, "License plate detection and recognition in unconstrained scenarios," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, Proceedings, Part XII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11216. Springer, 2018, pp. 593–609.

[37] S. Chen, C. Yang, J. Ma, F. Chen, and X. Yin, "Simultaneous end-to-end vehicle and license plate detection with multi-branch attention neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3686–3695, 2020.

[38] Y. Lee, J. Jeon, Y. Ko, M. Jeon, and W. Pedrycz, "License plate detection via information maximization," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14 908–14 921, 2022.

[39] S. Chen, S. Tian, J. Ma, Q. Liu, C. Yang, F. Chen, and X. Yin, "End-to-end trainable network for degraded license plate detection via vehicle-plate relation mining," *Neurocomputing*, vol. 446, pp. 1–10, 2021.

[40] S. M. Silva and C. R. Jung, "A flexible approach for automatic license plate recognition in unconstrained scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5693–5703, 2022.

[41] X. Fan and W. Zhao, "Improving robustness of license plates automatic recognition in natural scenes," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2022.

[42] C. L. P. Chen and B. Wang, "Random-positioned license plate recognition using hybrid broad learning system and convolutional networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 444–456, 2022.

[43] Y. Wang, Z. Bian, Y. Zhou, and L. Chau, "Rethinking and designing a high-performing automatic license plate recognition approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8868–8880, 2022.

[44] J. Pirgazi, M. M. Pourhashem Kallehbasti, and A. Ghanbari Sorkhi, "An end-to-end deep learning approach for plate recognition in intelligent transportation systems," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–13, 2022.

[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 9905. Springer, 2016, pp. 21–37.

[46] C. Zhang, Q. Wang, and X. Li, "V-LPDR: towards a unified framework for license plate detection, tracking, and recognition in real-world traffic videos," *Neurocomputing*, vol. 449, pp. 189–206, 2021.

[47] X. Lu, Y. Yuan, and Q. Wang, "AWFA-LPD: adaptive weight feature aggregation for multi-frame license plate detection," in *ICMR '21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021*, W. Cheng, M. S. Kankanhalli, M. Wang, W. Chu, J. Liu, and M. Worring, Eds. ACM, 2021, pp. 476–480.

[48] Q. Wang, X. Lu, C. Zhang, Y. Yuan, and X. Li, "Lsv-lp: Large-scale video-based license plate detection and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[49] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3733–3742.

[50] M. Ye, X. Zhang, P. C. Yuen, and S. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 6210–6219.

[51] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, ser. Lecture Notes in Computer Science, vol. 12356. Springer, 2020, pp. 776–794.

[52] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 9726–9735.

[53] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607.

[54] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *CoRR*, vol. abs/2003.04297, 2020.

[55] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[56] H. Hsu, T. Huang, G. Wang, J. Cai, Z. Lei, and J. Hwang, "Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 416–424.

[57] J. Gao, T. Han, Y. Yuan, and Q. Wang, "Domain-adaptive crowd counting via high-quality image translation and density reconstruction," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 8, pp. 4803–4815, 2023.

[58] G. Wang, Y. Wang, R. Gu, W. Hu, and J. Hwang, "Split and connect: A universal tracklet booster for multi-object tracking," *IEEE Trans. Multim.*, vol. 25, pp. 1256–1268, 2023.

[59] G. Wang, R. Gu, Z. Liu, W. Hu, M. Song, and J. Hwang, "Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 9856–9866.

[60] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[61] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 15 750–15 758.

[62] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[63] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[64] Y. Chen, X. Jin, X. Shen, and M. Yang, "Video salient object detection via contrastive features and attention modules," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, 2022, pp. 536–545.

[65] X. Wang, K. Zhao, R. Zhang, S. Ding, Y. Wang, and W. Shen, "Contrastmask: Contrastive learning to segment every thing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 11 594–11 603.

[66] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *CoRR*, vol. abs/1708.04552, 2017.

[67] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA*. IEEE Computer Society, 2017, pp. 936–944.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA*. IEEE Computer Society, 2016, pp. 770–778.

[69] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[70] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. D. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA*. Computer Vision Foundation / IEEE, 2019, pp. 658–666.

[71] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: an efficient and accurate scene text detector," in *2017 IEEE*

*Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2642–2651.

[72] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 379–387.

[73] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 408–417.

[74] S. Wang and H. Lee, "A cascade framework for a real-time statistical plate recognition system," *IEEE Trans. Inf. Forensics Secur.*, vol. 2, no. 2, pp. 267–282, 2007.

[75] H. Li, P. Wang, and C. Shen, "Towards end-to-end car license plates detection and recognition with deep neural networks," *CoRR*, vol. abs/1709.08828, 2017.

[76] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, and L. Huang, "Towards end-to-end license plate detection and recognition: A large dataset and baseline," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, vol. 11217. Springer, 2018, pp. 261–277.

[77] W. Zhang, Y. Mao, and Y. Han, "Slpnet: Towards end-to-end car license plate detection and recognition using lightweight CNN," in *Pattern Recognition and Computer Vision - Third Chinese Conference, PRCV 2020, Nanjing, China, October 16-18, 2020, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 12306. Springer, 2020, pp. 290–302.

[78] T.-A. Pham, "Effective deep neural networks for license plate detection and recognition," *The Visual Computer*, pp. 1–15, 2022.

[79] S. Montazzolli and C. R. Jung, "Real-time brazilian license plate detection and recognition using deep convolutional neural networks," in *30th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2017, Niterói, Brazil, October 17-20, 2017*. IEEE Computer Society, 2017, pp. 55–62.

[80] S. M. Silva and C. R. Jung, "Real-time license plate detection and recognition using deep convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 71, p. 102773, 2020.

[81] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada*, 2019, pp. 8024–8035.

**Haoxuan Ding** received the B.E. degree and the M.S. degree in aerospace propulsion theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018 and 2021 respectively. He is currently pursuing the Ph.D. degree from Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

**Junyu Gao** received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015 and 2021 respectively. He is currently an associate professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

**Yuan Yuan** (M'05-SM'09) is currently a Full Professor with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

**Qi Wang** (M'15-SM'15) 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 Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.