Boosting One-Stage License Plate Detector via Self-Constrained Contrastive Aggregation

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

***Abstract*—Scene Text Detection (STD) has applied in many fields successfully. One of the important applications of STD is License Plate Detection (LPD). As a unique identity of vehicle, License Plate (LP) facilitates the intelligent transportation in many fields, such as traffic enforcement, intelligent transportation dispatching, *etc*. However, there are many scene texts similar to LPs causing misjudgment of LP detector. To alleviate these disturbances, more discriminative features are necessary. In latent feature space, discriminative features should aggregate into a tight cluster to widen decision boundary. We assume three perspectives about how to aggregate features and boost feature expression. From these assumptions, a special contrastive triad is designed. Then, we propose a Self-Constrained Contrastive Aggregation (SCCA) method to lead the feature aggregation in latent space and boost the feature expression of backbone. The proposed SCCA is jointly trained with supervised learning for detection to improve the detection performance. The experiments show that our proposed SCCA prompts the baseline significantly and exceeds recent LP detectors, reaching 99.7 on both F1-score and AP on UFPR-ALPR dataset. Meanwhile, we compare the self-constrained contrastive learning with vanilla contrastive learning in experiments and visualize their LP features. The results show that our proposed SCCA reaches better performance and verifies our assumptions are reasonable.**

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

I. Introduction

With the development of deep learning, Scene Text Detection (STD) [1], [2] has acquired great success. One of the important applications of STD is License Plate Detection (LPD) in intelligent transportation system. As a unique identity of vehicle, the detection and recognition of License Plates (LPs) are beneficial to the traffic enforcement, intelligent transportation dispatching, *etc*. The Automatic License Plate Recognition (ALPR) system [3]–[8] has played a vital role in the development of smart city and intelligent traffic management. ALPR systems first detect the vehicles and LPs in images by License Plate Detection (LPD) module, and then feed the LP patches into License Plate Recognition (LPR) module. A high-performance detector for LPD is necessary to ensure the accuracy and robustness of downstream processing. Thanks to the feature extraction ability of Convolutional Neural Network (CNN), the detectors [9]–[11] learn to extract object-related features for LP detection by training, gaining promising performance.

Fig. 1. The disturbances from complex scene. The red boxes are ground truths. The blue boxes represent true predictions. The green boxes are false prediction causing by disturbances.

In general, the detectors learn how to distinguish foregrounds (*i.e.* LPs) from backgrounds (*i.e.* image contents) by supervised learning. The supervised learning minimizes the error between detection predictions and manual annotations (*i.e.* ground truths) by optimizing designed loss function. In order to minimize the loss function, the detectors learn to extract foregrounds-related features, achieving the detection of foregrounds. However, in complex street scene, there are abundant similar objects (Fig. 1) and they disturb the LP detectors. To alleviate the disturbances from backgrounds, the detectors need to extract more discriminative foreground features and widen the decision boundary between foregrounds and others.

To promote the specificity of objects and make foreground features discriminative, human beings exploit comparison and contrast to summarize the uniformity and difference among objects as prior knowledge for inference. To mimic the perception of the human, researchers propose the conception of contrastive learning [12]–[17]. Contrastive learning aims to make the features from the same source (*e.g.* same instance,

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.

same category, *etc.*) similar and make the features from different sources dissimilar. Specifically, a general strategy of contrastive learning in classification is as follows: first, a contrastive triad is designed which contains an original sample from dataset, an attractive sample with the category same as original, and a repulsive sample with the category different to original. Second, the neural network extracts features from the items in contrastive triad, and then the training of this neural network maximizes the similarity between original feature and attractive feature (*i.e.* pulls the original feature and attractive feature together) and minimizes the similarity between original feature and repulsive feature (*i.e.* pushes the original feature away from repulsive feature). Under this pull-and-push processing, the neural network learns what is the common characteristic between original and attractive and aggregates the features from the same category more tightly, extracting more discriminative features for classification.

Inspired by the contrastive learning, to extract more discriminative foreground features for LP detection, we need to design a strategy to squeeze and aggregate foreground features together tightly. We ask: *what are the keypoints to make foreground features for LP detection more aggregated?* We attempt to answer this question from following perspectives:

(i) Purpose of contrastive learning in LP detection: The contrast aims to pull or push the features of foreground objects into a tight cluster to widen the decision boundary in latent space for accurate detection. Comparing with the contrastive strategy in classification, the first thing is to build contrastive triad. Expect the original sample $\boldsymbol{I}$ with foreground objects from dataset, attractive sample $\boldsymbol{I}_+$ and repulsive sample $\boldsymbol{I}_-$ are necessary. The $\boldsymbol{I}_+$ pulls the $\boldsymbol{I}$ close to itself, and the $\boldsymbol{I}_-$ pushes the $\boldsymbol{I}$ away. Under this pull-and-push processing, the features of foreground objects (*i.e.* LPs) are squeezed into a tight cluster for discriminative expression.

(ii) Selection of feature distribution in LP detection with contrastive learning: We introduce the contrastive learning in detection to distinguish the foregrounds from backgrounds. However, the training of detection via supervised learning also aims at highlighting the feature expression of foregrounds. How to effectively prompt the performance of contrast needs to be considered. In joint training of contrsative learning and supervised learning, there are two typical distributions in latent space shown in Fig. 2. In Fig. 2a, the original $\boldsymbol{I}$ is sandwiched between repulsive $\boldsymbol{I}_-$ and attractive $\boldsymbol{I}_+$. The contrast under this distribution is achieved easily when joint training with supervised learning, because the goals of contrastive learning and supervised learning are same, which means the potential of contrast is not fully exploited in this distribution (Fig. 2a). On the contrary, in another distribution (Fig. 2b), $\boldsymbol{I}_+$ is far from $\boldsymbol{I}$ and the $\boldsymbol{I}_-$ is existed between $\boldsymbol{I}$ and $\boldsymbol{I}_+$. In training, the $\boldsymbol{I}_+$ will not pull the $\boldsymbol{I}$ close easily, and the $\boldsymbol{I}_-$ will not push the $\boldsymbol{I}$ far away. The optimization achieves a challenging contrast and boosts the effectiveness of contrast, aggregating foreground features more tightly. Thus, this pull-and-push processing is self-constrained, and we choose the distribution in Fig. 2b in contrastive learning.

(iii) How to build contrastive triad in LP detection: To build up the distribution in Fig. 2b, the positive sample

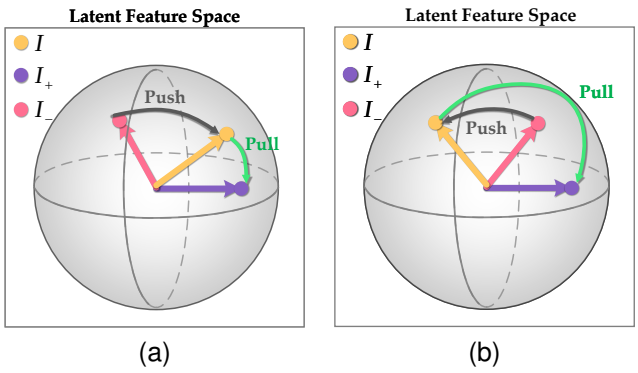


Fig. 2. Two distributions in contrastive learning.

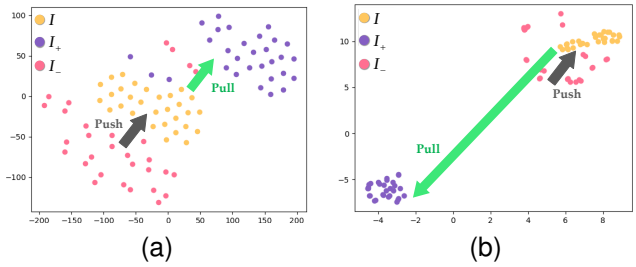


Fig. 3. The visualization results of LP features from distributions in Fig. 2. **(a)** the feature distribution described in Section IV-E. **(b)** the feature distribution described in Section III-A.

$\boldsymbol{I}_+$ and repulsive sample $\boldsymbol{I}_-$ have to be decided. $\boldsymbol{I}_+$ should be far away from $\boldsymbol{I}$, which means $\boldsymbol{I}_+$ is absolutely different to $\boldsymbol{I}$. Thus, we destroy the image content completely to get $\boldsymbol{I}_+$, ensuring the gap between them. And $\boldsymbol{I}_-$ should localize between $\boldsymbol{I}$ and $\boldsymbol{I}_+$ which means it should be similar to both $\boldsymbol{I}$ and $\boldsymbol{I}_+$. Therefore, we apply a weak augmentation on $\boldsymbol{I}$ and add $\boldsymbol{I}_+$ into it to gain $\boldsymbol{I}_-$. The contrastive learning in LP detection is based on this special contrastive triad.

We assume these three points to effectively introduce the contrastive learning into LP detection task. The first point claims that the contrast on contrastive triad will aggregate the foreground (*i.e.* LPs) features in latent space. The second point suggests that a challenging contrast will prompt the performance of contrastive learning in joint training with supervised learning, and according to this suggestion we select the feature distribution (Fig. 2b) for contrast. The third point shows how to build the specific contrastive triad for contrast discussed in the second point. To verify the correction of these three points, we visualize the feature distribution (Fig. 3) of traditional contrastive learning and the proposed method in this paper which are consistent in assumptions in Fig. 2.

Driven by this analysis, we propose a contrastive strategy in this paper, named Self-Constrained Contrastive Aggregation (SCCA), and introduce SCCA into LP detector to promote the detection accuracy. In SCCA, we first design a special contrastive triad different to vanilla contrastive learning. The attractive sample $\boldsymbol{I}_+$ is a content-ruined image instead of similar instance in vanilla contrastive learning, and the repulsive sample $\boldsymbol{I}_-$ is the weak augmentation similar to $\boldsymbol{I}$ rather than dissimilar instance in vanilla contrastive learning. In the joint

training of supervised learning and contrastive learning, the feature-level contrastive learning in SCCA aims to maximize the similarity between $\boldsymbol{I}$ and $\boldsymbol{I}_+$ and minimize the similarity between $\boldsymbol{I}$ and $\boldsymbol{I}_-$, and the supervised learning make the model focus on foregrounds to avoid the collapse of training. This joint training aggregates the original features into a tight cluster and extracts more discriminative features. The SCCA is jointly trained with supervised learning for detection to ensure the task-related features are learned. The SCCA is only used in training processing and there is no extra burden in inference.

In summary, the main contributions of this paper are:

1) Propose a contrastive learning strategy, Self-Constrained Contrastive Aggregation (SCCA), for LP detection task. The proposed SCCA method introduces self-supervision into LP detection task to make the backbone of detector extract more discriminative features. In joint training with supervised learning, the proposed SCCA promotes the detection performance effectively on three different LP detection datasets. To our best knowledge, this is the first time that contrastive learning is introduced into LP detection task.
2) To explore how to effectively exploit the contrast to aggregate features more tightly, we propose three assumptions about the feature distributions in contrast, and a special contrastive triad is designed according to our assumptions. The quantitative and visualization results all show that our assumptions in contrast are reasonable. The proposed SCCA method based on these assumptions boosts the performance of LP detectors effectively.
3) To prove the effectiveness of the proposed SCCA, SCCA is compared with vanilla contrastive learning method, and SCCA is better than vanilla contrastive learning. Ablation studies also explore and analyze the reason why SCCA is more efficacious.

The rest of this paper is organized as follows. Section II reviews the related works. Section III describes the proposed SCCA in details. Section IV illustrates the experimental results on LP detection datasets and analyzes the effect of our proposed SCCA method. Eventually, we summarize the whole work in Section V.

II. Related Work

In this section, the previous researches about object detection are briefly introduced in Section II-A. And then, the current deep learning based LP detection methods are summarized in Section II-B. In addition, the contrastive learning methods are exhibited in Section II-C.

A. Object Detection

The general object detectors mainly include two categories: one-stage detectors and two-stage detectors. Two-stage detectors [9], [18]–[21] first extract foreground object region proposals, and then classify the categories and regress the localization bounding boxes from the region proposals. However, these two cascade stages cost much time in detection, so the efficiency of two-stage detectors is low.

To improve the efficiency of detector, one-stage detectors remove the region proposal and directly classify the categories and regress bounding boxes from images. YOLO [10] and its variants [22]–[24] are popular one-stage detectors. Meanwhile, the implementation of multi-scale features improves the detection accuracy dramatically. The Single Shot MultiBox Detector (SSD) [25] uses multi-scale features to detect objects with different scales. Feature Pyramids Network (FPN) [26] also fuses the features from different layers to improve the detection performance. Because of the lack of region proposals, the extreme imbalance between foreground and background has a harmful effect on one-stage detectors. This is also the reason why two-stage detectors are more accurate than one-stage detectors. Thus, RetinaNet [27] introduces focal loss to focus on the training of hard sample detection, ignoring the easy negative backgrounds and boosting the detection accuracy of one-stage detectors dramatically. Furthermore, there are some methods breakthrough the traditional detection pipeline. The CornerNet [28] and CenterNet [29] convert the detection to keypoints localization. FCOS [11] first introduces the fully convolutional network (FCN) [30] into object detection. In summary, the imbalance of foregrounds and backgrounds is still the main limitation which hinders the improvement of one-stage detectors introduced above.

Recently, the transformer [31] has been introduced into detection task. DETR [32] is the first detection transformer which treats the detection as set prediction processing, removing the post-processing and realizing pure end-to-end framework. And some works [33]–[35] focus on accelerating the convergence of DETR [32]. However, LP is quite small in image and DETR is not suitable for small object detection [32], and the training of DETR needs extremely large datasets [36] which is still inadequate in LP detection task. Therefore, we choose a CNN-based method in this work as baseline rather than a transformer-based method.

B. LP Detection

Many object detection methods introduced above are implemented in LP detection task. Some researchers use two-stage detectors for LP detection. Dong *et al.* [37] use Faster R-CNN [9] in LP detection. Meanwhile, Rafique *et al.* [38] apply the R-CNN with exemplar-SVM to detect LPs. Li *et al.* [39] also build a Faster R-CNN [9] method to detect LPs, and a joint LP recognition module is combined into LP detector. However, two-stage LP detectors have more processing, causing low efficiency. Thus, two-stage detectors are not suitable for real-time LP detection.

Due to high efficiency, YOLO [10] and its variants [22]–[24] are widely used in LP detection task [40]–[43]. These YOLO-based methods directly introduce YOLO detectors into LP detection and they are lack of effective changes that aim at the characteristic of LPs. Towards LP detection in the wild, Silva and Jung [44], [45] propose the WPOD-NET and IWPOD-NET to detect LPs in unconstrained scenarios. To further improve the LP detection performance, Chen *et al.* [46] introduce multi-branch attention into end-to-end LP detectors. Chen and Wang [47] use semantic segmentation method to

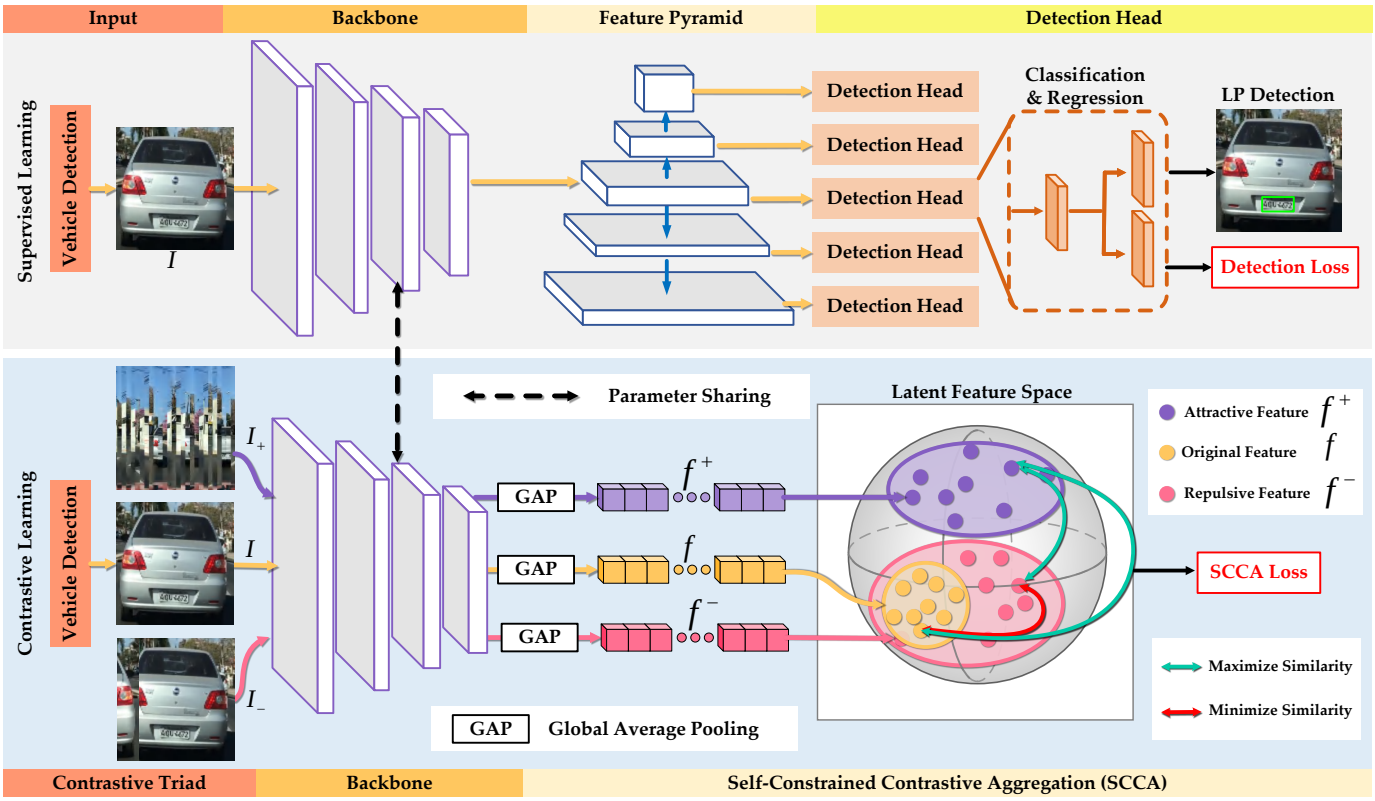


Fig. 4. The pipeline of the proposed SCCA method.

detect the LPs. In addition, some recent works [48], [49] focus on the detection of four corner vertex of LPs rather than the rectangle bounding boxes. These methods for open-world LP detection mainly focus on the localization methods for LP, and the similar scene texts are not considered deeply.

To alleviate the influence from similar scene texts, only Lee *et al.* [50] introduce a two-stream architecture in LP detection to detect LP and scene texts separately, and remove the non-LP texts in LP detection results. This work uses the conception of contrast to filter the non-LP texts in scene, but this is not a complete contrastive learning method. The recent contrastive learning methods are still not introduced into LP detection task to boost detection performance. Thus, we introduce the contrastive learning into LP detection task to alleviate the disturbance from similar scene texts

C. Contrastive Learning

Contrastive learning is an approach that concentrates on making the representations from similar instances closer and pushing away the representations from dissimilar instances. Most contrastive learning methods introduce the classification task in contrast. Bojanowski *et al.* [12] train the model to learn how to align the visual representations in low dimensional space. Wu *et al.* [13] utilize an instance-level classifier to distinguish the similar instances, achieving unsupervised image classification. Some contrastive learning methods under unsupervised scenario have prompted the downstream task performances comparable to supervised methods, such as MoCo [14], SimCLR [15], SwAV [16], BYOL [17], *etc.*

The contrastive learning methods learn to maximize the similarity in similar pairs and minimize the similarity in dissimilar pairs by optimizing contrastive loss. A general contrastive loss in face recognition is triplet loss [51], [52], which considers the the similar sample as an anchor to pull the original sample close and push the dissimilar sample away. N-Pair loss [53] contrasts the representations among one similar sample and $n-1$ dissimilar samples rather than only one similar and one dissimilar in triplet loss. Recently, inspired by Noise Contrastive Estimation (NCE), InfoNCE [54] is proposed and have become the most popular contrastive loss, which maximizes the similarity and minimizes the difference between the query representation with same category and other representations with different categories.

The contrastive learning method described above all need the distribution in Fig. 2a which can not exploits full potential on decoupling of foregrounds and backgrounds. Therefore, we design a special contrastive triad in this paper and propose a contrastive strategy which is different from contrastive learning introduced above.

III. Approach

In this section, we first explain how to build the contrastive triad in Section III-A. Then, Section III-B shows the pipeline of the proposed detector with SCCA. The details of SCCA are explained in Section III-C. Next, we analyze the reason why SCCA is effective in Section III-D. Finally, Section III-E explains other details of the training.

A. Contrastive Triad

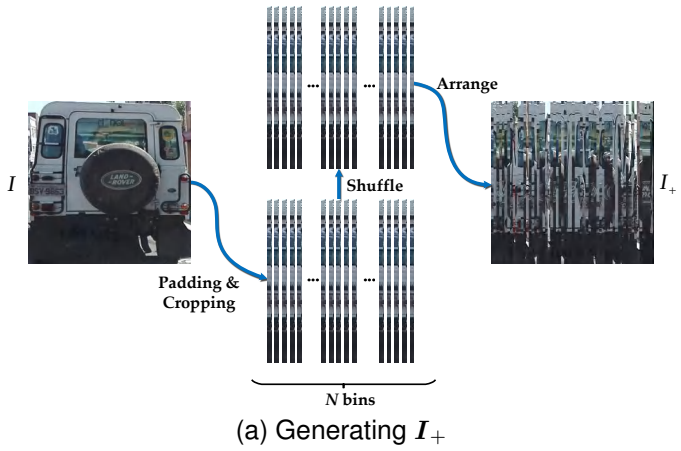


Fig. 5. The constitution of contrastive triad.

The contrastive learning on contrastive triad prompts the specificity of features. As shown in Fig. 2, there are two distributions, and we prefer to build the distribution in Fig. 2b to avoid the shortcut solution.

Under this circumstance, the attractive sample $\boldsymbol{I}_+$ should be far away from $\boldsymbol{I}$. This means the $\boldsymbol{I}_+$ needs to be different to $\boldsymbol{I}$ to keep the large gap between them. Meanwhile, the backbone extracts features from $\boldsymbol{I}$ and $\boldsymbol{I}_+$, and if there are LPs in both $\boldsymbol{I}$ and $\boldsymbol{I}_+$, the features of $\boldsymbol{I}$ and $\boldsymbol{I}_+$ should be similar because they all have LP-related features. However, in our expected distribution (Fig. 2b), the $\boldsymbol{I}_+$ should be dissimilar to $\boldsymbol{I}$, and this discrepancy confuses the backbone. Therefore, we destroy the image content by a rearrangement strategy in Fig. 5a to make sure the absolute difference between $\boldsymbol{I}$ and $\boldsymbol{I}_+$. $\boldsymbol{I}$ is divided into N bins and these bins are shuffled to rearrange a new image as $\boldsymbol{I}_+$.

The repulsive sample $\boldsymbol{I}_-$ should be close to original sample $\boldsymbol{I}$ and be located between $\boldsymbol{I}$ and $\boldsymbol{I}_+$. This means the distance between $\boldsymbol{I}_-$ and $\boldsymbol{I}_+$ needs to be smaller than that between $\boldsymbol{I}$ and $\boldsymbol{I}_+$. Meanwhile, the LP detection task is only a simple two-categories classification (foreground and background) task and the localization accuracy is more important for downstream application, so the noise (*i.e.* the translation of LP position) on LP position should be introduced into contrast to teach LP detectors to localize LPs accurately. Thus, we introduce a standard flip-and-shift strategy [55] to add noise into $\boldsymbol{I}$. And then, to guarantee the distance between $\boldsymbol{I}_-$ and $\boldsymbol{I}_+$ is smaller than that between $\boldsymbol{I}$ and $\boldsymbol{I}_+$. The transformed image is added with $\boldsymbol{I}_+$. The generation of repulsive sample $\boldsymbol{I}_-$ is shown in Fig. 5b. The generation of $\boldsymbol{I}_-$ is defined as Eq. 1:

$$\begin{aligned} \widetilde{\boldsymbol{I}} &= \mathcal{T}(\boldsymbol{I}), \\ \boldsymbol{I}_- &= \alpha\widetilde{\boldsymbol{I}} + (1-\alpha)\boldsymbol{I}_+, \end{aligned} \tag{1}$$

where $\mathcal{T}$ is the standard flip-and-shift strategy [55], and α is a weighting parameter.

The original image $\boldsymbol{I}$, the attractive image $\boldsymbol{I}_+$, and the repulsive image $\boldsymbol{I}_-$ constitute the contrastive triad $\mathcal{C}$ in proposed SCCA method.

B. Pipeline

Considering the efficiency in application under traffic scenario, we select one-stage detector for LP detection in this paper. The whole pipeline is shown in Fig. 4. It contains two main parts, a supervised learning part for detection and a contrastive learning part for SCCA.

The supervised learning part is used to train the detector by traditional supervised learning strategy. In the supervised learning part, the image is first input to a pre-trained vehicle detector in [41] to get vehicle patches $\boldsymbol{I}$ (*i.e. original sample*). Then the $\boldsymbol{I}$ is input to a backbone, and the backbone extracts features from $\boldsymbol{I}$. These multi-scale features are fed into FPN [26] to fuse the low-level features and high-level features. Finally, a shared detection head predicts the bounding boxes on the features from FPN [26].

In contrastive learning part, the vehicle patch $\boldsymbol{I}$ from vehicle detector is first used to generate $\boldsymbol{I}_+$ and $\boldsymbol{I}_-$ according to the strategies in Section III-A to obtain contrastive triad $\mathcal{C}$. Then, $\boldsymbol{I}$, $\boldsymbol{I}_+$, and $\boldsymbol{I}_-$ are input the shared backbone with supervised learning part to extract features for contrastive learning. Those features are fed into SCCA. The details of SCCA are described in Section III-C.

C. Self-Constrained Contrastive Aggregation (SCCA)

As shown in Fig. 4, the proposed SCCA is based on the feature-level contrastive learning. The items in $\mathcal{C} = \{\boldsymbol{I}, \boldsymbol{I}_+, \boldsymbol{I}_-\}$ are fed into backbone $\mathcal{F}$ to gain features for contrast. To reduce the computation, a global average pooling layer is utilized to condense the latent features into embedding vectors, denoted by $\boldsymbol{f}$, $\boldsymbol{f}^+$, and $\boldsymbol{f}^-$ respectively:

$$\begin{aligned} \boldsymbol{f} &= GAP(\mathcal{F}(\boldsymbol{I})), \\ \boldsymbol{f}^+ &= GAP(\mathcal{F}(\boldsymbol{I}_+)), \\ \boldsymbol{f}^- &= GAP(\mathcal{F}(\boldsymbol{I}_-)), \end{aligned} \tag{2}$$

where $GAP(\cdot)$ represents the global average pooling along with the spatial dimension. The contrastive learning is exerted on these three embedding vectors.

In proposed SCCA, the backbone needs to learn how to pull the $\boldsymbol{f}$ close to $\boldsymbol{f}^+$ and push the $\boldsymbol{f}$ away from $\boldsymbol{f}^-$ simultaneously. This purpose is illustrated in Eq. 3:

$$\begin{aligned} &\max Sim(\boldsymbol{f}, \boldsymbol{f}^+), \\ &\min Sim(\boldsymbol{f}, \boldsymbol{f}^-), \\ &\min Sim(\boldsymbol{f}^-, \boldsymbol{f}^+), \\ s.t.\ &Sim(\boldsymbol{f}^-, \boldsymbol{f}^+) > Sim(\boldsymbol{f}, \boldsymbol{f}^+), \end{aligned} \tag{3}$$

where $Sim(\cdot, \cdot)$ represents the quantitative similarity between inputs, *e.g.* L_2 distance, cosine similarity, *etc.* In this paper, the $Sim(\cdot, \cdot)$ is cosine similarity (Eq. 4), a standard metric for similarity measure:

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|}, \tag{4}$$

where $\cos(\boldsymbol{a}, \boldsymbol{b})$ is the cosine similarity between $\boldsymbol{a}$ and $\boldsymbol{b}$. The constrained condition in Eq. 3 is used to make the $\boldsymbol{f}^-$ localize at the position between $\boldsymbol{f}$ and $\boldsymbol{f}^+$. This means the $\boldsymbol{f}^-$ likes a constraint boundary to prevent the cross of $\boldsymbol{f}$. This is because the attractive feature $\boldsymbol{f}^+$ from content-ruined image could be considered as noise. If the $\boldsymbol{f}$ crosses the $\boldsymbol{f}^-$ and aggregates with $\boldsymbol{f}^+$, the training of detection will collapse. Thus, we use this constrained condition and the joint training of supervised learning to avoid the training collapse.

In SCCA, we design a contrastive loss to optimize and achieve Eq. 3. Specifically, $\boldsymbol{f}_i$, $\boldsymbol{f}_i^+$, $\boldsymbol{f}_i^-$ are the embedding vectors from i-th layer features. The cosine similarity gauges their similarity, and the loss function for contrastive learning is shown in Eq. 5:

$$\begin{aligned} l_i^{ctr} = &|1 - \cos(\boldsymbol{f}_i, \boldsymbol{f}_i^+)| + \\ &|\cos(\boldsymbol{f}_i, \boldsymbol{f}_i^-)| + |1 - \cos(\boldsymbol{f}_i^-, \boldsymbol{f}_i^+)|, \end{aligned} \tag{5}$$

where l_i^{ctr} is the contrastive loss for i-th layer in model.

In a hierarchical backbone, multi-scale features effectively facilitate the accurate detection, so that the multi-scale features are also introduced into the proposed SCCA method. To take ResNet [56] as an example, the outputs from the 2-nd, 3-th, and 4-th layers of ResNet (denoted as $Conv2\_x$, $Conv3\_x$, and $Conv4\_x$ in [56]) are fed into SCCA for contrastive learning. The SCCA loss is defined as follows:

$$\mathcal{L}_{SCCA} = \frac{1}{3}(\lambda_1 l_2^{ctr} + \lambda_2 l_3^{ctr} + \lambda_3 l_4^{ctr}), \tag{6}$$

where $\mathcal{L}_{SCCA}$ is the SCCA loss. l_2^{ctr}, l_3^{ctr}, and l_4^{ctr} are the contrastive losses for the output features from 2-nd, 3-th, and 4-th layers in ResNet [56]. The λ_1, λ_2, and λ_3 are corresponding weighting coefficients.

D. Effect Analysis

As demonstrated in Fig. 6, there is a huge gap between $\boldsymbol{f}$ and $\boldsymbol{f}^+$. Meanwhile, the $\boldsymbol{f}_-$ is located at this gap. In training of SCCA, the changing process in latent feature space is demonstrated in Fig. 6. The supervised learning part keeps the gap between $\boldsymbol{f}$ and $\boldsymbol{f}^+$. And the training of SCCA forces the $\boldsymbol{f}$ close to remote $\boldsymbol{f}^+$, and makes $\boldsymbol{f}$ move close to $\boldsymbol{f}^-$. But the surrounded $\boldsymbol{f}^-$ pushes the $\boldsymbol{f}$ away from $\boldsymbol{f}^+$. In detail, the maximization of similarity between $\boldsymbol{f}$ and $\boldsymbol{f}^+$ will pull

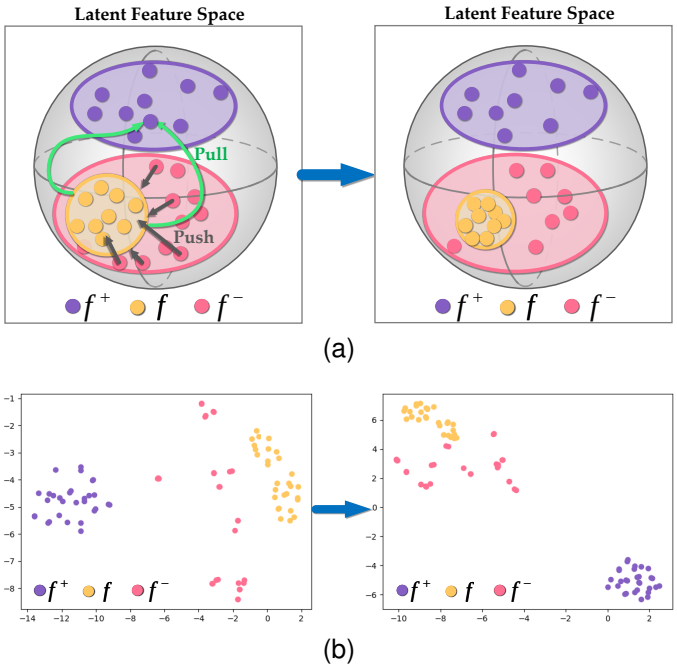


Fig. 6. SCCA achieves feature aggregation in latent space.

the $\boldsymbol{f}$ along with the direction to $\boldsymbol{f}^+$, but because of the supervised learning, the $\boldsymbol{f}$ can only reach the feature $\boldsymbol{f}_-$ instead of the $\boldsymbol{f}^+$ ($\boldsymbol{f}^+$ is considered as a noise in detection). Meanwhile, the minimization of similarity between $\boldsymbol{f}$ and $\boldsymbol{f}^-$ will arrest $\boldsymbol{f}$ for moving to $\boldsymbol{f}^+$ and push the $\boldsymbol{f}$ far away from $\boldsymbol{f}^+$. In this adversarial pull-and-push processing, the pulling from $\boldsymbol{f}^+$ and the pushing from $\boldsymbol{f}^-$ squeeze $\boldsymbol{f}$ more tightly in latent space, and the joint training of supervised learning ensures that the $\boldsymbol{f}$ focuses on foregrounds rather than destroyed content in $\boldsymbol{f}^+$ to avoid the training collapse of LP detection. Under this self-constrained contrastive learning, the distribution of $\boldsymbol{f}$ is squeezed into a tight cluster, achieving the feature aggregation. To verify the pull-and-push processing described above, we visualize the change of features in latent space by tSNE [57] in Fig. 6b. From Fig. 6b, we find that the training of SCCA indeed aggregates $\boldsymbol{f}$ obviously, which is consistent in the assumption in Fig. 6a. In addition, this feature aggregation phenomenon in latent space is also proved through visualization analysis in Section IV-E.

E. Other Details

As mentioned in Eq. 3, to keep the gap between $\boldsymbol{f}$ and $\boldsymbol{f}^+$ and concentrate on the LP-related features, we train the detection task with SCCA jointly. For one-stage detectors, the detection loss $\mathcal{L}_{det}$ usually contains three core parts, *i.e.* classification loss $\mathcal{L}_{cls}$, regression loss $\mathcal{L}_{reg}$, and other special losses $\mathcal{L}_{other}$. In joint training, considering the detection loss, the total loss function $\mathcal{L}_{total}$ is defined by Eq. 7:

$$\begin{aligned} \mathcal{L}_{det} &= \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{other}, \\ \mathcal{L}_{total} &= \mathcal{L}_{det} + \eta \mathcal{L}_{SCCA}, \end{aligned} \tag{7}$$

where $\mathcal{L}_{cls}$ is the cross-entropy loss for classification, and $\mathcal{L}_{reg}$ is the regression loss which is GIoU loss [58]. The $\mathcal{L}_{other}$ denotes auxiliary loss functions, such as centerness loss

(a) UFPR-ALPR

(b) CCPD-Base

(c) SSIG-SegPlate

Fig. 7. The detection results from detectors with SCCA on three datasets. The **red boxes** are ground truths, and **green boxes** are predicted bounding boxes.

in FCOS [11]. The $\mathcal{L}_{SCCA}$ represents the loss for proposed SCCA, illustrating in Eq. 6. The η is a scale factor to adjust the weights of $\mathcal{L}_{det}$ and $\mathcal{L}_{SCCA}$. The value of η is 0.1 to ensure $\mathcal{L}_{det}$ and $\mathcal{L}_{SCCA}$ have the same magnitude.

IV. Experiments

This section demonstrates the comparison results of detector with proposed SCCA. Section IV-A illustrates the evaluation metrics for LP detection task. Section IV-B introduces three LP detection datasets used in this paper. Section IV-C explains the experimental details. Section IV-D demonstrates the comparative performance of proposed SCCA. Eventually, Section IV-E further analyzes the influence of the proposed SCCA.

A. Evaluation Metrics

To evaluate the detection performance, the general evaluation metrics in detection are illustrated in Eq. 8:

$$\begin{aligned} P &= \frac{TP}{TP+FP}, \\ R &= \frac{TP}{TP+FN}, \\ IoU(\boldsymbol{pred}, \boldsymbol{gt}) &= \frac{\boldsymbol{pred} \cap \boldsymbol{gt}}{\boldsymbol{pred} \cup \boldsymbol{gt}}, \\ F1\text{-}score &= \frac{2 \times P \times R}{P+R}, \\ AP &= \int_0^1 p(r)dr, \end{aligned} \tag{8}$$

where TP, FP, and FN denote the True Positive, False Positive, and False Negative predictions respectively. P, R, and IoU denote the Precision, Recall, and Intersection over Union in detection. In this paper, the IoU threshold τ used to decide TP, FP, and FN is 0.5 in testing. The F1-score and Average Precision (AP) are used to comprehensively consider the precision (P) and recall (R). In detection, the precision and recall are contradictory. Thus, researchers draw the P-R curve $p(r)$ based on detection results to express the relation between precision and recall, and consider the trade-off between them according to $p(r)$. AP is the area under P-R curve $p(r)$. The detectors with higher F1-score and AP perform better.

B. Datasets

In this paper, three popular LP detection datasets are utilized to evaluate the performance of detectors.

UFPR-ALPR: UFPR-ALPR dataset [41] contains 150 videos ($4,500$ images) with LP bounding boxes shot by on-board camera from moving vehicles. Every frame only contains a single annotated Brazilian LP. The frames have a fixed size of 1920×1080 pixels. And there are three types of LP: gray LP, red LP, and motorcycle LP. The training set and testing set both include 60 videos ($1,800$ images), and the validation set has 30 videos (900 images).

CCPD: Chinese City Parking Dataset (CCPD) [59] is a large scale Chinese LP dataset. It contains $250,000$ images (with the size of 720×1160) shot by handheld devices in

TABLE I
THE COMPARISON RESULTS FOR DETECTOR WITH THE PROPOSED SCCA METHOD. THE DETECTOR WITH PROPOSED SCCA METHOD ACHIEVES BETTER PERFORMACNE COMPARED TO OTHER LP DETECTORS.

Method	Dataset	Backbone	P	R	F1-score	AP	FPS
FCOS (Baseline) [11]	UFPR-ALPR	ResNet-18	92.3	94.3	93.3	88.2	38.6
FCOS-SCCA (Ours)		ResNet-18	99.6	99.8	**99.7** (+6.4)	**99.7** (+11.5)	38.6
YOLOv2 [22]	UFPR-ALPR	Darknet-19	92.7	94.7	93.7	87.8	111.4
YOLOv3 [23]		Darknet-53	94.9	97.4	96.1	-	47.6
YOLOv4 [24]		CSPDarknet-53	93.1	92.6	92.8	88.5	19.8
EAST [61]		PVANet	92.8	99.9	96.2	-	38.5
R-FCN [62]		ResNet-101	94.6	99.8	94.5	-	15.1
Laroca *et al.* [41]		FAST-YOLO	-	98.3	-	-	66.5
FGFA [63]		ResNet-50	97.2	98.3	97.7	-	11.5
FCOS (Baseline) [11]	CCPD-Base	ResNet-18	99.3	100.0	99.6	99.4	35.6
FCOS-SCCA (Ours)		ResNet-18	99.6	100.0	**99.8** (+0.2)	**99.7** (+0.3)	35.6
Cascade Classifier [64]	CCPD-Base	-	55.4	-	-	47.2	32.0
SSD300 [25]		VGG-16	99.1	-	-	94.4	40.0
YOLOv2 [22]		Darknet-19	98.8	-	-	93.1	42.0
YOLOv4-Tiny [24]		CSPDarknet-53-Tiny	94.7	-	-	90.2	110.2
YOLOv4 [24]		CSPDarknet-53	99.1	-	-	98.8	33.1
Faster R-CNN [9]		VGG-16	98.1	-	-	92.9	15.0
TE2E [65]		VGG-16	98.5	-	-	94.2	3.0
RPnet [59]		-	99.3	-	-	94.5	61.0
SLPNet [66]		ShuffleNetv2	99.9	99.3	99.6	-	25.0
Silva and Jung [67]		Darknet-19	86.1	-	-	-	13.6
Pham [68]		-	99.3	-	-	98.1	168.3
FCOS (Baseline) [11]	SSIG-SegPlate	ResNet-18	94.7	100.0	97.2	95.3	37.5
FCOS-SCCA (Ours)		ResNet-18	95.8	99.6	**97.6** (+0.4)	**96.1**(+0.8)	37.5
Silva and Jung [43]	SSIG-SegPlate	FAST-YOLO	95.1	99.5	97.3	-	8.7
Laroca *et al.* [41]		FAST-YOLO	-	100.0	-	-	-
Silva and Jung [69]		FAST-YOLO	95.1	99.5	97.3	-	8.7
Laroca *et al.* [42]		Fast-YOLOv2	95.3	99.8	97.5	-	29.4

parking lot. Each image includes a single LP instance. There are nine subset: 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). We use 50% data in CCPD-Base for training and other 50% data in CCPD-Base for testing.

SSIG-SegPlate: SSIG-SegPlate dataset [60] has 2,000 images for Brazilian LP detection with the size of 1920×1080 pixels. There are 1,762 images for passenger vehicles, 118 images for buses or trucks, and 120 images for motorcycles. SSIG-SegPlate dataset contains 800 images for training, 800 images for testing, and 400 images for validation.

C. Experimental Details

The SCCA method in this paper is proposed for one-stage detectors, so we select frequently-used one-stage detector, FCOS [11], to verify the effect of SCCA method.

The entire image is first fed into a vehicle detection network used in [41], and when the confidence threshold is set as 0.125, the recall of vehicle is 100% to ensure all of the vehicles are detected successfully. The vehicle patches in dataset is resized to 512×512 as original sample $\boldsymbol{I}$. Then, $\boldsymbol{I}$ is used to generate attractive sample $\boldsymbol{I}_+$ and repulsive sample $\boldsymbol{I}_-$ (Section III-A). In training, the LP detector is trained 30 epochs. The initial learning rate is 10^{-3}, and the learning rate is set to $1/10\times$ every 10 epochs.

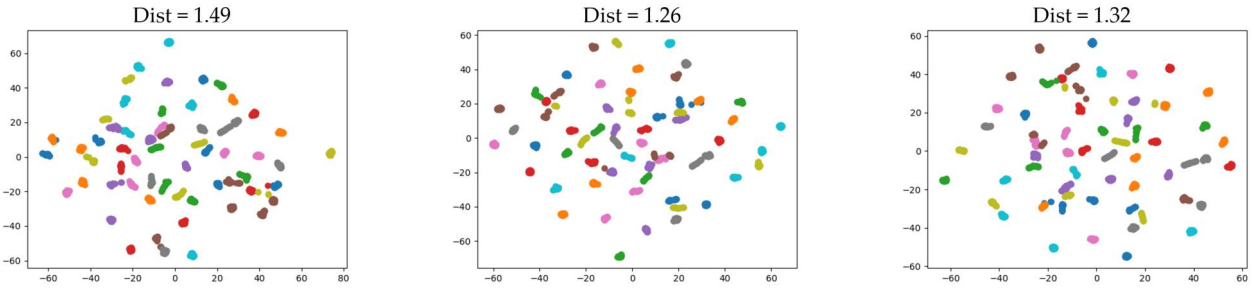
The experiments are implemented on the equipment with Intel(R) CPU Core(TM) i7-6900K @ 3.4GHz, 64GB RAM, and a single NVIDIA GTX 1080TI GPU. The method is built by Pytorch framework [70].

D. Performance

*1) **Comparison on LP detetcors**:* We first compare the proposed method with recent LP detectors. Fig. 7 demonstrates some results on used datasets. Table I summarizes the comparative results among recent LP detectors on UFPR-ALPR, CCPD-Base, and SSIG-SegPlate datasets.

UFPR-ALPR: On UFPR-ALPR dataset, the proposed SCCA promotes the performance of FCOS (baseline) significantly. There is a distinct increase of 6.4(6.8%) on F1-score and 11.5(13.0%) on AP. Meanwhile, the FCOS with SCCA achieves the best detection accuracy in compared LPD methods, reaching 99.7 on F1-score and 99.7 on AP.

CCPD-Base: On CCPD-Base dataset, although the FCOS (baseline) has already achieved high performance, our SCCA still improves the performance of baseline, which means the introduction of SCCA helps the detector to process hard samples in dataset. The FCOS with SCCA achieves 99.8 on F1-score and 99.7 on AP, which is the highest among recent LPD. Compared to other LP detectors, FCOS with SCCA reaches extremely high value on AP.



(a) Baseline (b) Ours (c) Vanilla

Fig. 8. The visualizations of LP features through tSNE [57]. The Dist on top is the average intra-cluster distance.

SSIG-SegPlate: On SSIG-SegPlate, the FCOS (baseline) also reaches promising performance, and only the hard samples need to be tackled. Although the increase of performance is small, the hard sample detection also benefits from the proposed SCCA. The SCCA brings an increase of 0.4 on F1-score and 0.8 on AP, making the FCOS with SCCA obtain the best performance in comparison.

In summary, the proposed SCCA method prompts the detection accuracy of baseline detector effectively and achieves the best performance in comparison with the *state-of-the-art* in recent years.

*2) **Comparison on recent object detetcors**:* Some advanced object detectors are proposed, but these methods are still not used in LP detection task. Thus, to evaluate the progress of proposed method, we compare our proposed SCCA with recent general object detectors on the most challenging dataset in Section IV-B (*i.e.* UFPR-ALPR dataset [41]). Table IIdemonstrates the comparison results of detection accuracy among recent general object detectors.

TABLE II
THE COMPARISON RESULTS ON UFPR-ALPR [41] AMONG RECENT GENERAL OBJECT DETECTORS.

Method	Venue	P	R	AP	FPS
FCOS-SCCA (Ours)	-	99.6	99.8	**99.7**	38.6
YOLOF† [71]	CVPR'21	-	-	**99.7**	22.1
YOLOX† [72]	ArXiv'21	-	-	98.6	55.1
VarifocalNet† [73]	CVPR'21	-	-	99.5	12.3
Sparse R-CNN† [74]	CVPR'21	-	-	96.6	12.9
TOOD† [75]	ICCV'21	-	-	99.5	12.6
DDOD† [76]	ACM MM'21	-	-	99.6	12.0
Scaled YOLOv4 [77]	CVPR'21	95.2	100.0	99.4	47.4
ObjectBox [78]	ECCV'22	96.4	89.7	94.7	34.6

† The compared methods are reproduced through open source object detection toolbox MMdetection [79].

From Table II, we find that our proposed SCCA and YOLOF [71] achieves the best detection performance (99.7 AP), and our proposed SCCA method has higher efficiency with 38.6 FPS compared to 22.1 FPS of YOLOF [71]. Although YOLOX [72] and Scaled YOLOv4 [77] have higher inference speed than our proposed method, their detection accuracy is lower than SCCA, and our SCCA method also gains 38.6 FPS in inference which is suitable for real-time LP detection. The comparison results shows that our proposed SCCA has great potential on LP detection task. In addition, there are some works about detection transformer, such as DETR [32], Deformable DETR [33]. However, detection transformer is not suitable for small object detection [32], and the detection transformer can not converge on small dataset [36] (UFPR-ALPR dataset [41] only includes 1,800 images), so that the detection transformer performance is extremely weak in LP detection. Thus, we do not introduce detection transformer in comparison.

E. Ablation Studies

To further explore and analyze the effectiveness of the proposed SCCA, some ablation studies are implemented on UFPR-ALPR dataset [41].

TABLE III
THE COMPARISON RESULTS ON UFPR-ALPR [41] FOR TWO DISTRIBUTIONS IN CONTRAST.

Method	P	R	F1-score	AP
FCOS (Baseline) [11]	92.3	94.3	93.3	88.2
FCOS-SCCA (Ours)	99.6	99.8	**99.7** (+6.4)	**99.7** (+11.5)
FCOS-Vanilla	97.6	99.4	98.5	97.1

Comparison between two distributions: In Fig. 2, we assume that there are two possible distributions in latent space during contrastive learning. And we exploit the self-constrained distribution in Fig. 2b to further improve the performance. To verify this hypothesis, we build up another contrastive strategy (*i.e.* vanilla contrastive learning). In vanilla contrastive learning, we introduce the color jittering and Gaussian blur in $\boldsymbol{I}$ to generate $\boldsymbol{I}_+$. There is no translation in $\boldsymbol{I}_+$, so that the $\boldsymbol{I}_+$ is extremely similar to $\boldsymbol{I}$. And the standard flip-and-shift strategy [55] in Section III-A is applied on $\boldsymbol{I}$ to generate $\boldsymbol{I}_-$. The comparison results are illustrated in Table III. The FCOS with proposed SCCA achieves the better performance, and exceeds the accuracy of vanilla contrastive learning. Although the detection accuracy benefits from the vanilla contrastive learning, the SCCA boosts the accuracy more significantly.

Aggregation Analysis: To directly prove the advantage of proposed SCCA, the LP features are cropped from whole

image features from 2-nd layer of RenNet [56]. The tSNE [57] is utilized to visualize the LP features into a two-dimension space. As shown in Fig. 8, the points with the same color represent a single LP instance from frames in the same video. Fig. 8a demonstrates the features from baseline (FCOS [11]). Fig. 8b and Fig. 8c are features from SCCA and vanilla contrastive learning. The average intra-cluster distance is shown on the top to quantitatively measure the feature aggregation ability. In Fig. 8a, the points in a cluster are scattered (Dist=1.49) compared to Fig. 8b and Fig. 8c. The points from SCCA and vanilla contrastive learning are more dense in clusters compared to baseline, and the proposed SCCA achieves better aggregation performance (Dist=1.26) than vanilla contrastive learning (Dist=1.32). The visualization demonstrates that our three assumptions for feature aggregation in Section I are reasonable. The self-constrained contrastive learning aggregates the feature tighter than vanilla contrastive learning and boosts the backbone feature expression effectively.

TABLE IV
THE COMPARISON RESULTS ON UFPR-ALPR [41] TO FIND THE BEST λ_i IN EQ. 6

λ_1	λ_2	λ_3	P	R	F1-score	AP
1	0	0	99.6	99.8	**99.7**	**99.7**
0	1	0	97.7	99.9	98.8	97.6
0	0	1	98.1	99.2	98.6	97.4
1	1	0	97.8	99.9	98.8	99.5
1	0	1	97.5	99.0	98.3	96.6
0	1	1	96.7	99.4	98.0	96.9
1	1	1	98.3	99.8	99.0	98.9
2	1	1	99.0	99.7	99.3	98.9
1	2	1	98.3	99.6	98.9	97.9
1	1	2	97.6	99.0	98.3	96.7

Comparison of λ_i: In Eq. 6, multi-scale features are introduce into SCCA to fully exploit the contrast among features from different layers. We attempt to adjust λ_i to explore the importance of multi-scale features. Table IV illustrates the results under different configurations of λ_i. When only the low-level features is used, the detector with SCCA achieves highest accuracy. This is because the LPs are small in vehicle patches, and the low-level features with high-resolution are more suitable for small object detection. Thus, the low-level features facilitate the LP detection task most. Meanwhile, the contrastive learning on low-level features could be propagated to high-level features because of the utilization of FPN [26]. Thus, the comparison results show that we should only use the features from 2-nd layer in ResNet [56].

TABLE V
THE COMPARISON RESULTS ON UFPR-ALPR [41] TO FIND THE BEST η IN EQ. 6

η	P	R	F1-score	AP
0.01	96.6	98.8	97.7	95.5
0.10	99.6	99.8	**99.7**	**99.7**
1.00	97.4	97.8	97.6	96.3

Comparison of η: To balance the detection loss $\mathcal{L}_{det}$ and SCCA loss $\mathcal{L}_{SCCA}$, there is a hyper-parameter η to adjust the weights of $\mathcal{L}_{det}$ and $\mathcal{L}_{SCCA}$ in $\mathcal{L}_{total}$. Table V shows the experimental results in comparison of η. When the $\eta = 0.01$, the weight of $\mathcal{L}_{SCCA}$ is tiny, and the training of detector focuses on the optimization of $\mathcal{L}_{det}$. The potential of proposed SCCA could not be exposed obviously. When the η increases to 1, the detection performance decreases rapidly. This is because the self-constrained contrast in SCCA keeps the gaps among $\boldsymbol{I}$, $\boldsymbol{I}_+$, and $\boldsymbol{I}_-$ all the time. Therefore, the $\mathcal{L}_{SCCA}$ could not be optimized to 0. In training, the detection loss $\mathcal{L}_{det}$ reduces and reaches a small value when the detection performance raises. Under this condition, a large $\mathcal{L}_{SCCA}$ will mislead the optimization and guide the detector to an optimization dilemma, causing the low detection accuracy.

TABLE VI
THE COMPARISON RESULTS ON UFPR-ALPR [41] TO FIND THE BEST α IN EQ. 1

α	P	R	F1-score	AP
0.9	99.6	99.8	**99.7**	**99.7**
0.99	95.8	99.3	97.5	96.7
0.999	94.5	99.2	96.8	95.3

Comparison of α: In Eq. 1, to make sure the repulsive sample $\boldsymbol{I}_-$ is closer to $\boldsymbol{I}_+$ than $\boldsymbol{I}$ (Fig. 2b), the $\boldsymbol{I}_-$ should contain some information from $\boldsymbol{I}_+$. Therefore, we add the $\boldsymbol{I}_+$ into $\boldsymbol{I}_-$ by Eq. 1, and several combination ratio α configurations are considered in Table VI. With the increase of α, the detection accuracy decreases. This may because the increase of α enlarges the distance between $\boldsymbol{I}_-$ and $\boldsymbol{I}_+$, corrupting the construction of ideal distribution in Fig. 2b. Finally, the selected configuration in our SCAA is $\alpha = 0.9$.

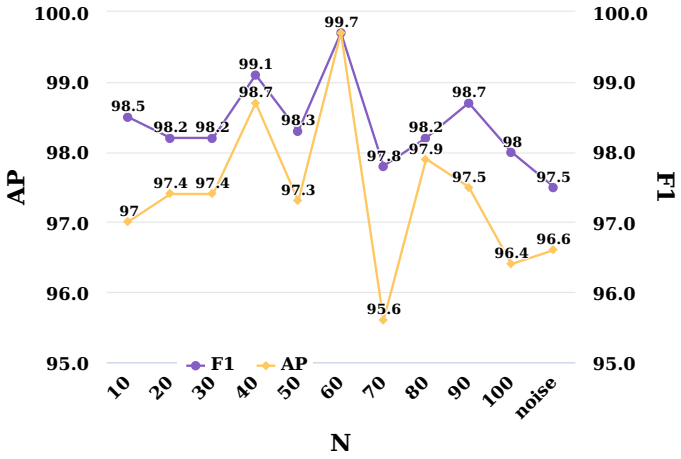


Fig. 9. The influence about the N in generating of $\boldsymbol{I}_+$. The experiments are implemented on UFPR-ALPR [41].

Comparison of N: In Section III-A, the original sample $\boldsymbol{I}$ is splits to N column bins, and these bins are rearrange to generate $\boldsymbol{I}_+$. From Fig. 9, we find that with the increase of N, the detection accuracy first increases and then decreases, and reaches the highest with $N = 60$. This may because when the N is less than 60, the LP information (*i.e.* some existed parts of LP in image) is not damaged totally, and the information leak disturbs the localization in contrast. When

the N increases lager than 60, the whole image information is destroyed and does harm to the understanding of scenes. To verify this assumption, we directly take a noise as $\boldsymbol{I}_+$ and find that the accuracy is low (*i.e.* the noise at horizontal axis in Fig. 9). Thus, $N = 60$ is the best configuration in experiments.

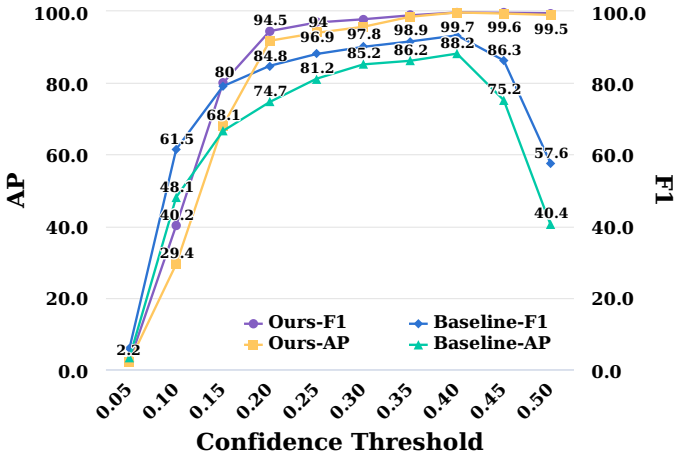


Fig. 10. The influence about the confidence score threshold in inference. The experiments are implemented on UFPR-ALPR [41].

The influence of confidence threshold in inference: In inference, the LP detector filters the predictions by confidence to remove the distinct false positive predictions. To comparison the prediction quality of detectors, we increase the confidence score gradually, the comparison results are demonstrated in Fig. 10. With the increase of confidence threshold, the detection accuracy increases, because the predictions with low confidence are removed. However, when the confidence threshold is larger than 0.4, the accuracy of baseline (*i.e.* FCOS) decreases dramatically. This means the baseline predicts many true positive predictions with the confidence scores around 0.4. In the same condition, our SCCA still maintains a high accuracy, which means the detector with SCCA extracts more discriminative features for detection and predicts the true positive samples with high confidence.

V. Conclusion

This paper presents a contrastive learning method, Self-Constrained Contrastive Aggregation (SCCA), for LP detection. The SCCA improves the LP detection accuracy of one-stage LP detector significantly. Specifically, we first assume three keypoints to aggregate the features and design a contrastive triad according to these assumptions. And then, SCCA method is designed to achieve the self-constrained contrastive learning on this special triad. The proposed SCCA is introduced into the training for LP detection task, boosting the feature expression of backbone and prompting the detection performance effectively. The experimental results show that the detector with SCCA has higher accuracy than the baseline and other LP detectors. The ablation studies verify our assumptions about how to aggregate feature are reasonable and show the details which influence the detection accuracy dramatically. Meanwhile, the SCCA method is only used in training rather than inference, which means the SCCA method will not bring any extra burdens in inference, maintaining the efficiency of one-stage detector.

In this paper, contrastive learning is only implemented on the original images and corresponding transformation images. To further expend the proposed SCCA into general object detection, we will introduce more extra contrastive strategies into SCCA method in the future, such as contrast among categories, contrast among video frames, *etc.*

References

[1] T. Guan, C. Gu, C. Lu, J. Tu, Q. Feng, K. Wu, and X. Guan, "Industrial scene text detection with refined feature-attentive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6073–6085, 2022.

[2] Y. Cai, C. Liu, P. Cheng, D. Du, L. Zhang, W. Wang, and Q. Ye, "Scale-residual learning network for scene text detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2725–2738, 2021.

[3] R. K. Srinivas, P. Shivakumara, H. A. Jalab, R. W. Ibrahim, G. H. Kumar, U. Pal, and T. Lu, "Riesz fractional based model for enhancing license plate detection and recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2276–2288, 2018.

[4] S. Du, M. Ibrahim, M. S. Shehata, and W. M. Badawy, "Automatic license plate recognition (ALPR): A state-of-the-art review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 311–325, 2013.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[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] 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.

[12] P. Bojanowski and A. Joulin, "Unsupervised learning by predicting noise," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 517–526.

[13] 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.

[14] 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.

[15] 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, 2020, pp. 1597–1607.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: delving into high-quality region proposal network with adaptive convolution," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada*, 2019, pp. 1430–1440.

[22] 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.

[23] ——, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] W. Wang, J. Zhang, Y. Cao, Y. Shen, and D. Tao, "Towards data-efficient detection transformers," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, ser. Lecture Notes in Computer Science, vol. 13669. Springer, 2022, pp. 88–105.

[37] 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.

[38] 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.

[39] H. Li, P. Wang, and C. Shen, "Towards end-to-end car license plates detection and recognition with deep neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1126–1136, 2017.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] ——, "A flexible approach for automatic license plate recognition in unconstrained scenarios," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2021.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.

[52] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823.

[53] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 1849–1857.

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

[55] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 3008–3017.

[56] 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.
[57] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] 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.
[63] 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.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] T.-A. Pham, "Effective deep neural networks for license plate detection and recognition," *The Visual Computer*, pp. 1–15, 2022.
[69] 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.
[70] 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.
[71] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 13 039–13 048.
[72] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: exceeding YOLO series in 2021," *CoRR*, vol. abs/2107.08430, 2021.
[73] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, "Varifocalnet: An iou-aware dense object detector," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 8514–8523.
[74] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: end-to-end object detection with learnable proposals," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 14 454–14 463.
[75] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 3490–3499.
[76] Z. Chen, C. Yang, Q. Li, F. Zhao, Z. Zha, and F. Wu, "Disentangle your dense object detector," in *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. César, F. Metze, and B. Prabhakaran, Eds. ACM, 2021, pp. 4939–4948.
[77] C. Wang, A. Bochkovskiy, and H. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 13 029–13 038.
[78] M. Zand, A. Etemad, and M. A. Greenspan, "Objectbox: From centers to boxes for anchor-free object detection," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13670. Springer, 2022, pp. 390–406.
[79] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *CoRR*, vol. abs/1906.07155, 2019.

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.