Vehicle-Mounted Adaptive Traffic Sign Detector for Small-Sized Signs in Multiple Working Conditions

Junfan Wang, *Student Member, IEEE*, Yi Chen, *Student Member, IEEE*, Xiaoyue Ji, *Member, IEEE*, Zhekang Dong, *Senior Member, IEEE*, Mingyu Gao, and Chun Sing Lai, *Senior Member, IEEE*

***Abstract*—Traffic sign detection is of great significance to the development of the Intelligent Transportation System (ITS) as a database for environmental awareness. The main challenges of existing traffic sign detection method are inaccurate small object detection, difficult mobile deployment, and complex working environment. Based on these, a vehicle-mounted adaptive traffic sign detector (VATSD) for small-sized signs in multiple working conditions is proposed in this paper. First, the Backbone of the detector is optimized. A feature tight fusion structure is designed to constitute a new feature extraction module, DCSP, which improves the feature extraction capability and the detection accuracy of small objects with negligible additional parameters. Second, an image enhancement network IENet with an adaptive joint filtering strategy is proposed. The IENet enables the dynamic selection of filters and thus adaptively optimizes low-quality images under multiple conditions to improve the accuracy of subsequent detection tasks. The proposed method has experimented on three traffic sign datasets and the detection accuracy increased by up to 7.6% compared to the original. The proposed detector demonstrates superiority over other state-of-the-art (SOTA) methods in terms of small object detection accuracy, detection speed, and environmental adaptability. Further, we deployed VATSD to Jetson Xavier NX and achieved a detection speed of 21.6 FPS, meeting real-time requirements.**

***Index Terms*—Adaptive joint filtering, image enhancement, small objects, traffic sign detection**

I. INTRODUCTION

Intelligent Transportation System (ITS) can understand and sense road conditions to reduce traffic accidents [1]. Traffic sign detection (TSD), as a crucial sub-module, enables drivers or intelligent vehicles to make timely decisions to control and improve driving safety by quickly and accurately conveying traffic information [2]. Excellent TSD methods can advance autonomous driving by integrating into Autonomous Driving Systems (ADS) or Advanced Driver Assistance Systems (ADAS).

Long-range TSD can provide sufficient response time for the driver or the autonomous vehicle, but it requires a model with high feature extraction capability for small-sized targets. Considering that ADS and ADAS require fast and accurate detection of traffic signs, general two-stage networks [3-5] such as Faster RCNN cannot be deployed on the vehicle side due to their large computing and memory costs. The general one-stage networks [6, 7] sacrifice part of the detection accuracy to improve the detection speed, especially for the poor detection performance of small-sized objects. Existing researches improve the feature extraction ability of the detector through the attention mechanism [8, 9] or by combining localization, segmentation and other modules [2, 10, 11]. For example, Shen *et al.* [8] designed a group multi-scale attention (GMSA) module that aggregates features in the foreground and ignores irrelevant information, which can enable the network to focus more on small object regions and improve the detection accuracy of small traffic signs. Min *et al.* [2] proposed a multiscale densely connected object detector relying on an effective feature fusion strategy to improve the detection of small traffic signs. However, the extra modules will slow down the detection speed of the detector and increase its memory cost for in-vehicle deployment.

Based on the above, many researchers have improved their feature extraction capabilities by improving the feature extraction module of the network itself [12-15], among which the optimization of lightweight networks [16-20] is a good remedy. Lightweight networks such as CSPNet [18], MobileNet [19] and VGG [21] accomplish traffic sign tasks with their simple and efficient backbones, but simple use of them will affect the detection accuracy of small-sized objects due to their limited model width and depth. Some researches [22-24] improve the detection ability of small objects based on the improvement of lightweight backbone. However, most of the above methods are general object detection methods. The traffic sign detector deals with traffic signs with multiple features, consistent morphology, and diverse features, which is different from general targets with multi-scale variation and large feature variance. It is difficult for the general object detection network to efficiently complete the feature extraction of traffic signs, and it is easy to extract background features by mistake and lose some traffic sign features at the same time.

This work was supported in part by the Key R&D Project of Hangzhou under grants No. 2022AIZD0009, 2022AIZD0022, the Key Research and Development Program of Zhejiang Province grant No. 2022C01062 *(Corresponding author: Zhekang Dong).*

J. Wang and M. Gao are with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou, China, 310018, and also with the Zhejiang Provincial Key Lab of Equipment Electronics, Hangzhou, China, 310018, (e-mail: wangjunfan@hdu.edu.cn, mackgao@hdu.edu.cn).

X. Ji and Y. Chen are with the College of Electrical Engineering, Zhejiang University, Hangzhou, China, 310027, (e-mail: ji.xiaoyue@zju.edu.cn, morningone@126.com).

Z. Dong is with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou, China, 310018, and also with the College of Electrical Engineering, Zhejiang University, Hangzhou, China, 310027, (e-mail: englishp@hdu.edu.cn).

C. S. Lai is with the Department of Electronic and Computer Engineering, Brunel University London, London, UB8 3PH, UK and also with the School of Automation, Guangdong University of Technology, Guangzhou, China 510006 (email: chunsing.lai@brunel.ac.uk)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TITS.2023.3309644, IEEE Transactions on Intelligent Transportation Systems

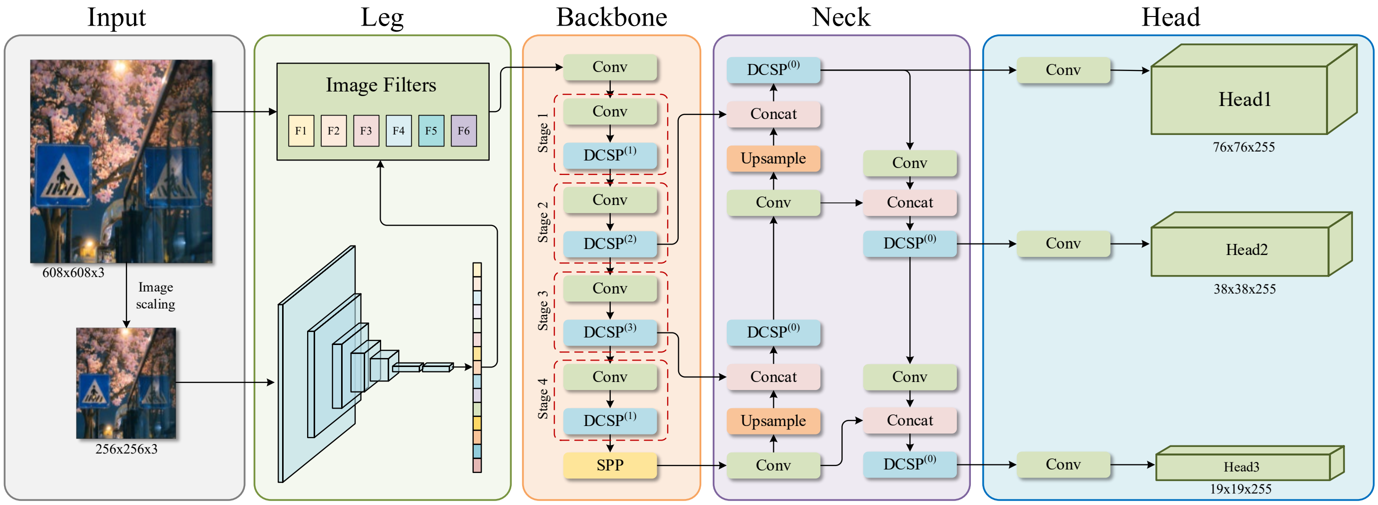


Fig. 1. Architecture of VATSD. x in DCSP$^{(x)}$ represents the number of resnet blocks contained in the DCSP.

This paper considers part of the feature extraction process for traffic signs assigned using dynamic convolutions. Dynamic convolution is able to simultaneously extract multi-faceted features of traffic signs (e.g., color, shape, content details, etc.) and allows more diverse features to be extracted. Based on this, our group designed a two-stage feature tight fusion structure, by fusing the multi-faceted features of dynamic convolution, and then fusing residuals to further extract traffic sign features.

On the other hand, TSD needs to be applicable to complex environments, such as dimly lit tunnels and foggy weather, which may lead to poor-quality captured images and increase the difficulty of object detection. Existing studies always optimized these low-quality images by image enhancement. Zheng *et al* [25] introduced a generative adversarial network with an expanded feature pyramid generator to improve the detection accuracy of images with motion blur. Liang *et al*. [26] achieved optimization of dim images through a Recurrent Exposure Generation (REG) module and seamlessly connected it with Multi-Exposure Detection (MED) to suppress non-uniform illumination and noise problems. Yu *et al*. [16] proposed a fusion model based on YOLOv3 and VGG19 networks to achieve accurate detection of traffic signs underexposure and dimness using relationships in multiple images. Yuan *et al*. [27] proposed a color angle model to provide color differentiation information as a way to improve the detection of traffic signs after color changes. However, [25, 26] only considered traffic sign detection under a single condition. Although [16, 27] realizes the processing of multi-working conditions, they operate the same for all working conditions, and cannot adaptively process visual information according to changes in light and color like the brain.

In this paper, a vehicle-mounted adaptive traffic sign detector (VATSD) for small-sized signs in multiple working conditions is proposed. It not only provides a better trade-off between detection accuracy and detection speed but also considers the effect of complex conditions on detection results. First, a feature tight fusion structure in the Backbone of VATSD is designed, and the proposed dynamic cross stage partial (DCSP) module with negligible additional parameters can improve the feature learning capability and the detection accuracy of small-sized traffic signs. Secondly, an image enhancement network IENet based on an adaptive joint filtering strategy is proposed to achieve real-time optimization of low-quality images by the dynamic combination of multiple filters, thus improving the subsequent detection accuracy. Finally, the proposed detector is deployed on Jetson Xavier NX for practical road tests to verify its good practical significance.

Our main contributions in the present work can be summarized as follows:

- A novel feature tight fusion structure is designed to improve the Backbone of the detector. The proposed feature extraction module DCSP improves the feature extraction capability for small traffic signs with negligible additional parameters, achieving a trade-off between detection accuracy and detection speed.
- An image enhancement network IENet with an adaptive joint filtering strategy is proposed. IENet achieves fast and efficient optimization of dynamic scenes under complex working conditions, thus improving the accuracy of subsequent object detection.
- A vehicle-mounted adaptive traffic sign detector (VATSD) for small-sized signs in multiple working conditions is proposed. The proposed method is evaluated on three traffic sign datasets and deployed on Jetson Xavier NX to illustrate the superior small object detection capability and excellent model performance.

II. VATSD Architecture

A vehicle-mounted adaptive traffic sign detector suitable for multiple complex conditions is proposed in this paper. For the complexity of the driving environment, the adaptive filtering strategy is proposed thus the designed IENet accomplishes adaptive optimization of the images. Second, the proposed DCSP module enriches the network structure with negligible parameters to enhance the network feature extraction capability. The structure of the object detector is shown in Fig. 1.

The network structure is divided into five parts: Input, Leg, Backbone, Neck, and Head. This paper is mainly for the optimization of Leg and Backbone part. These two components optimize the input images and extract features from them, ensuring the reliability of the subsequent detection results. IENet is proposed as the Leg of the network. The image is input to IENet after size scaling, and the image quality is improved through the adaptive joint filtering strategy, to improve the subsequent detection effect. In Backbone, a feature tight fusion structure is designed to fuse CSPNet and dynamic convolution [28] and DCSP is proposed. The existing feature fusion structures suffer from the feature information disparity issue, resulting in the unavailability of certain fused feature information. In this paper, we leverage a two-stage process of gradually fusing shallow features with deep features to achieve tight fusion of features. The specific tight fusion method will be described in detail in Section III, which can improve the representation of small objects.

III. Dynamic Cross Stage Partial Module

The proposed feature tight fusion structure combines CSPNet and dynamic convolution, which can improve the feature learning ability for small-size traffic signs. Similar with CSP, DCSP based on the feature tight fusion structure can be applied to a variety of network structures such as ResNet [29] and DenseNet [30] to improve the network structure. Jocher *et al*. [23]optimized the structure of CSPNet on the backbone of YOLOv5 to improve the performance of the detector. In order to illustrate the effectiveness of the DCSP structure, this paper compares the performance of DCSPResNet, CSPResNet and YOLO-CSPResNet based on ResNet [31-33].

The DCSP first divides the input according to the channel dimension: $x_0 = [x_0', x_0'', x_0''']$. First, x_0'' uses dynamic convolution to improve the feature extraction ability, and then merges it with x_0' and implements the Transition operation [18]. Transition operation represents the transition layer of the module, which mainly contains 1×1 convolutional layer and pooling layer. x_0''' goes through Transition, ResNet, and Transition respectively. Finally, the above three parts are merged through the Concat operation. The proposed DCSP is shown in Fig. 2(d). It can be seen that DCSP performs two Concat and Transition operations on the divided three-part input to ensure that the features under a single path are preserved to the greatest extent and that the feature information of the three paths is fully integrated.

Since the standard dynamic convolution will bring additional computational cost and parameter cost, the dynamic convolution decomposition (DCD) is used to solve this limitation. A low-parameter dynamic convolution is designed, which aims to improve the feature extraction ability with negligible additional parameters. The specific expression of dynamic convolution is shown in (1)

$$W_d = \sum_{k=1}^{K} \pi_k(x) W_k$$
$$s.t. 0 \le \pi_k(x) \le 1, \sum_{k=1}^{K} \pi_k(x) = 1 \tag{1}$$

where K represents the number of convolution combinations, W_k represents the parameter of the k^{th} ordinary convolution, $\pi_k(x)$ represents the attention parameter given for the input x, and W_d represents the parameter of the dynamic convolution.

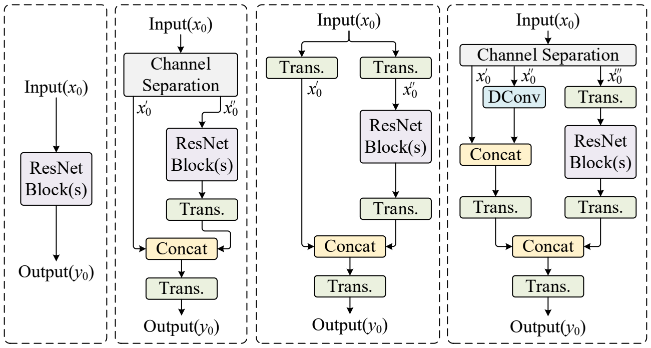


Fig. 2. Various structures in which the CSP module is applied to ResNet. (a) ResNet [29], (b) CSP-ResNet [18], (c) YOLOCSP-ResNet [31], and (d) proposed DCSP-ResNet. Trans. represents the Transition operation and DConv represents the dynamic convolution layer.

The dynamic convolution is a combination of K ordinary convolutions, the attention parameters need to be computed for each convolution, so the number of parameters will increase by a factor of K compared to the ordinary convolution. To find a balance between model size and performance, the DCD is invoked to optimize the dynamic convolution in DCSP.

Assuming that the weight of each ordinary convolution can be decomposed into a static weight W_0 and an offset weight ΔW_k, W_0 represents the average of the weights of all ordinary convolutions, and ΔW_k represents the difference between the ordinary convolution weight and the static weight. The dynamic convolution can be rewritten by (2).

$$W_d = \sum_{k=1}^{K} \pi_k(x)\left(W_0 + \Delta W_k\right) = W_0 + \sum_{k=1}^{K} \pi_k(x)\Delta W_k, \quad W_0 = \sum_{k=1}^{K} W_k \tag{2}$$

where the main computational cost is on the right-hand side. Assuming that the channel of the ordinary convolution is C, after decomposing the singular value decomposition (SVD) on the right, it can be seen that the calculated hidden channel is enlarged to KC, which is the fundamental reason for a large amount of calculation of the dynamic convolution.

The dynamic channel fusion mechanism is used to address the limitations of dynamic convolution, which is implemented using a full matrix $\Phi(x)$, where each element $\phi_{i,j}(x)$ is a function of the input x. $\Phi(x)$ is an $L \times L$ matrix, where $L << C$. The key idea is to significantly reduce the dimensionality in the latent space to achieve a more compact model. Dynamic convolution is implemented with dynamic channel fusion using

$$W_d = W_0 + P\Phi(x)Q^T = W_0 + \sum_{i=1}^{L}\sum_{j=1}^{L} p_i \phi_{i,j}(x) q_j^T$$
$$s.t. P \in \mathbb{R}^{C\times L}, \Phi(x) \in \mathbb{R}^{L\times L}, Q \in \mathbb{R}^{C\times L} \tag{3}$$

where Q is a $C \times L$ matrix, which is used to compress the input to a lower channel space and the number of channels is compressed from C to L. The result is dynamically fused with $\Phi(x)$. Finally, the channel is re-expanded to the original C channel through the P matrix. The dynamic convolution constructed in this way can greatly reduce the computational cost without affecting its performance.

Take the input of a matrix with 256 channels and all random values as an example, $x_0 \in \mathbb{R}^{(40,40,256)}$. The MAC and parameter

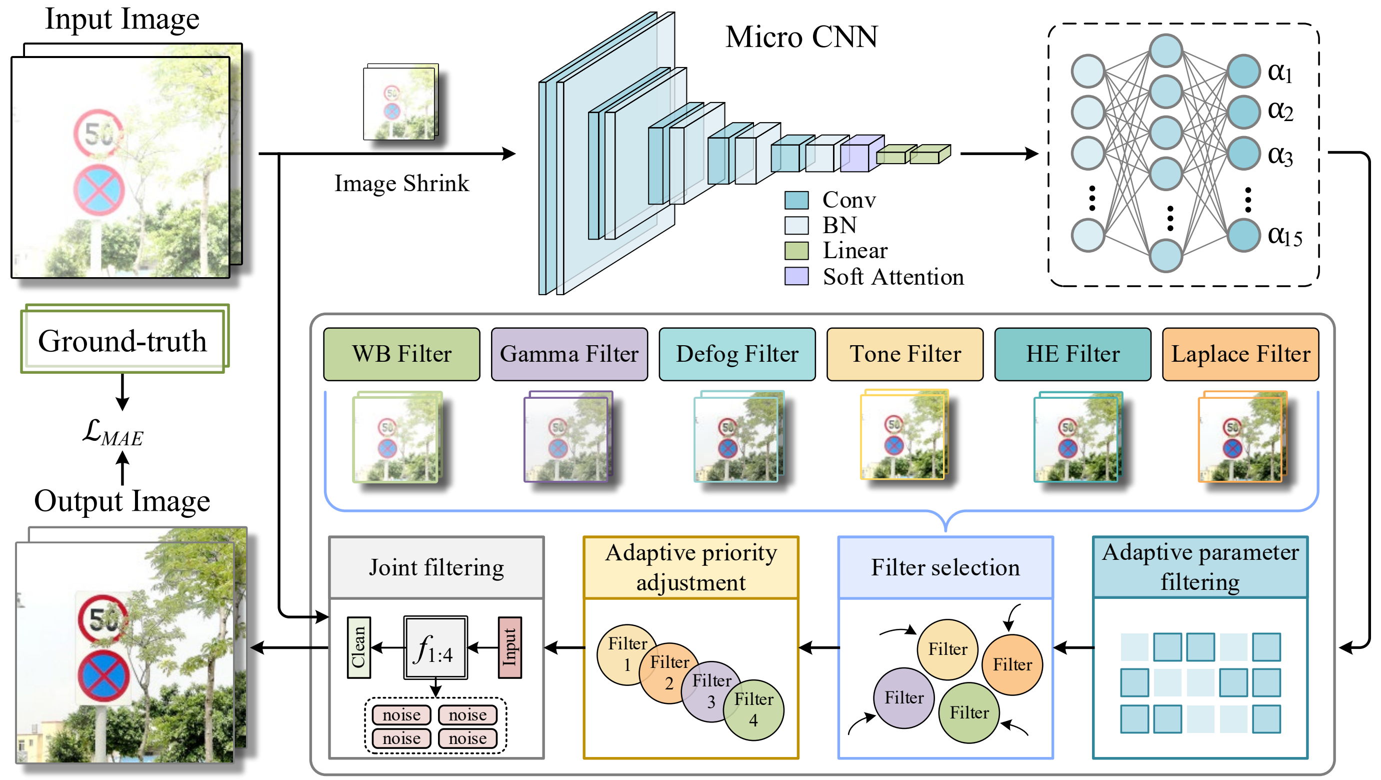


Fig. 3. Structure of IENet. α_1-α_{15} are the required parameters for the six filters, which are trained by the small neural network Micro CNN.

TABLE I
MAC AND PARAMETERS OF DIFFERENT CSP STRUCTURES

Net.	MAC	Params.
CSP-ResNet	0.396G	0.247
YOLOCSP-ResNet	0.501G	0.313
DCSP-ResNet	0.463G	0.426
DCSP-ResNet with DCD	0.422G	0.263

quantities under different CSP structures are calculated, and the results are shown in Table I. Multiply-accumulate (MAC) operations used to evaluate the computational intensity and efficiency of the network. The MAC of the DCSP without DCD is fewer compared to YOLOCSP-ResNet, but greater compared to CSP-ResNet. And the number of parameters is larger than these two methods. The results demonstrate that the competitiveness of DCSP cannot be adequately showcased without the inclusion of DCD. The DCSP optimized by DCD has a certain decrease in MAC and parameter amount, and the parameter amount is reduced by half compared with that before the optimization, which satisfies the trade-off between detection speed and detection.

IV. ADAPTIVE IMAGE ENHANCEMENT NETWORK

Based on the working conditions of existing traffic sign detectors, the following five conditions are selected for image optimization: fog/haze, exposure, blur, fading, and dimness. These five working conditions involve the processing of lighting, color, resolution, etc., and basically cover the image problems existing in existing driving scenes. Considering the memory and computing cost, this paper selects six filters: white balance filter, gamma filter, Defoe filter, hue filter, histogram equalization filter and Laplacian filter, based on adaptive joint filtering strategies to achieve targeted handling of multi-working-condition driving scenarios.

The proposed adaptive joint filtering strategy is similar to the human visual perception mechanism. The human visual system [34] is able to process images of different scenes using a combination of various mechanisms and strategies, enabling people to perceive and interpret visual information in various environments. Likewise, an adaptive joint filtering strategy can adjust its processing strategy according to specific characteristics of the image, such as illumination, color, or resolution. As shown in Fig. 3, the training of Micro CNN can well simulate the learning process of the brain. After the learning is completed, it can be evaluated under different working conditions. Similar to the selective attention of the brain, visual information is selectively processed and prioritized according to the perceived relevance and importance of stimuli. The adaptive joint filtering strategy can also select the optimal filter combination strategy to complete the image processing through the evaluation score of the image. This process can not only achieve targeted optimization of multi-working-condition driving scenarios, but also improve processing efficiency. The above-mentioned specific process is shown in the Algorithm 1. L1 loss [35] is used to optimize the neural network by calculating the mean absolute error (MAE) loss $\mathcal{L}_{MAE}$ of the clean image and the filtered image.

Next, the adaptive joint filtering strategy is described in detail. First, the soft attention in Micro CNN implements the quality assessment of the image, and its calculation is as follows:

$$att\left(\alpha,m\right)=\sum_{n=1}^{N}\alpha_i m_i \tag{4}$$

where α represents the soft attention weight, m represents the feature map, and the most important weight is selected by weighted summation. The soft attention weights as follows:

$$\alpha=softmax\left(z\right)=\frac{e^{z_i}}{\sum_j e^{z_j}} \quad z=avp\left(m\right) \tag{5}$$

where the $avp(\ )$ function represents the global average pooling, which can effectively reduce the dimensionality of m.

By adding a soft attention layer to the Micro CNN, the Micro CNN is able to adaptively adjust the required filter parameters and suppress the effects of other filters on the image for the complex working conditions of the input image. Here *softmax* is applied for feature normalization, and the normalized value is compared with the filter initialization threshold to realize adaptive joint filtering

The adaptive joint filtering strategy provides the optimal filter combination and filter order according to the current environment, achieving efficient image optimization. On the one hand, it can solve the additional parameter amount and calculation cost caused by repeated filtering, and on the other hand, it can effectively avoid the negative optimization of the image caused by the filter working independently.

The parameters of White Balance filter and Gamma filter are generated directly by Micro CNN as shown in (6) and (7).

$$I_{WB}\left(x,y\right)=\left(W_b B\left(x,y\right),W_g G\left(x,y\right),W_r R\left(x,y\right)\right) \tag{6}$$

$$I_{ga}\left(x,y\right)=I\left(x,y\right)^{\gamma} \tag{7}$$

where x, y represents the position of a pixel of the image, $I(x,y)$ represents the original image, $I_{WB}(x, y)$ and $I_{ga}(x, y)$ represent the image after the White Balance filter and Gamma filter respectively. $B(x, y)$, $G(x, y)$ and $R(x, y)$ represent the three color channels in the original image (blue, green, red). W_b, W_g, W_r represent the parameters needed for the White Balance filter and γ represents the parameters needed for the Gamma filter.

Based on the dark channel priori algorithm [36] and the atmospheric scattering model [37], (8) is used to represent the fogged image.

$$I\left(x\right)=J\left(x\right)t\left(x\right)+A\left(1-t\left(x\right)\right) \tag{8}$$

where $I(x)$ represents the original image, and $J(x)$ represents the target image. $t(x)$ is the transmission transmittance from the scene to the camera. A is global atmospheric light. To get a clear $J(x)$, the key is to get A and $t(x)$. To do this, we can directly pass $I(x)$ to A, and $t(x)$ is calculated as

$$t(x)=1-\min_{y\in\Omega(x)}(\min_c(I^c(y)/A^c)) \tag{9}$$

where the superscript c indicates the three channels of RGB, $\Omega(x)$ represents a window centered at pixel x.

To make the image more realistic, this paper corrects (9), by making Micro CNN adaptively learn a parameter ω. The corrected equation is shown below:

```
Algorithm 1 Adaptive joint filtering strategy
Input: Image with noise $I$;
Result: Image after joint filtering $I_c$;
Initialize Micro CNN networks $M$ and load model weights, filter parameter threshold $T$=0.5, $I_c = I$;
Image Shrink $I_s \in \mathbb{R}^{128\times128\times3} \leftarrow I$;
Output preliminary filter parameters $\mathcal{A}(\alpha_1, \alpha_2,\ldots, \alpha_{15})$ by forward $I_s$ in $M$, $\mathcal{A} \leftarrow M(I_s)$;
Normalize weights $\mathcal{B}(\beta_1, \beta_2,\ldots, \beta_{15}) \leftarrow \mathcal{A}$, where $\beta_j \in (0,1)$;
if $\beta_n > T$ then
  Retain weight;
else
  Delete weight;
Obtained by anti-normalization $\mathcal{C}(c_1, c_2, \ldots, c_n)$, $n$<=15;
According to weight $\mathcal{C}$ choice and initialize filters
Sort filter in order $c_{\max} \rightarrow c_{\min}$;
for $f$ in filters do
  $I_c = f(I_c)$;
end
return $I_{c;}$
```

$$t(x)=1-\omega\min_{y\in\Omega(x)}(\min_c(I^c(y)/A^c)) \tag{10}$$

Based on (10), a clear defogged image can be obtained as shown in (11):

$$J(x)=(I(x)-A)/\max(t(x),t_0)+A \tag{11}$$

where t_0 is the partiality factor that prevents the denominator of the expression from being zero, which we set to 0.01 in our experiments.

The Tone filter [38] is designed as a monotonic and segmented linear function. Micro CNN gets the points (k/8, T_k/T_8) on the tone curve by learning 8 parameters (t_0, t_1, …, t_7) and calculating the prefix and T_k of all parameters. The transformation of the Tone filter is shown in (12).

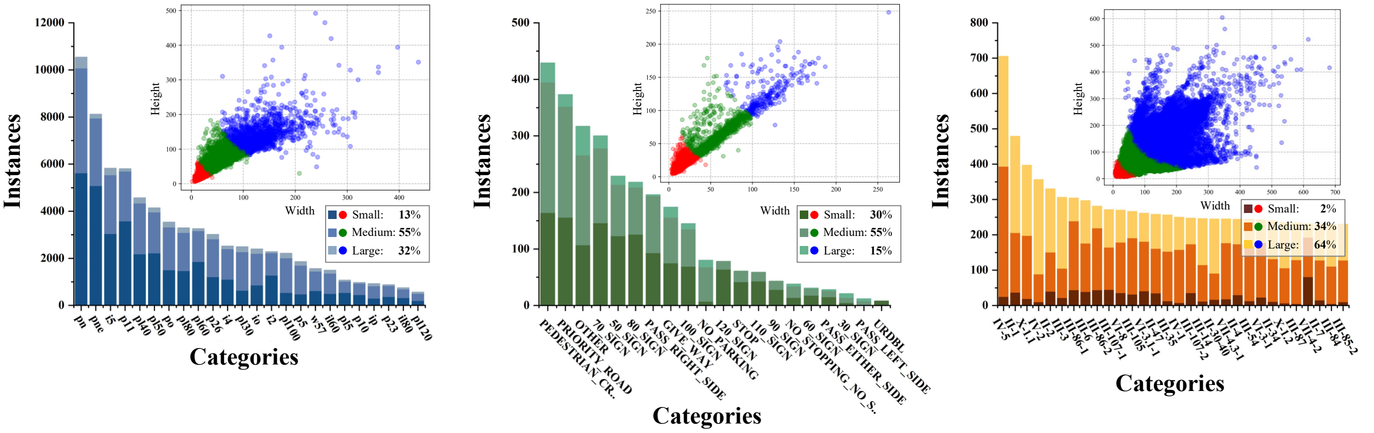
$$\begin{gathered} I_{tone}(x,y)=\frac{1}{T_8}\sum_{k=0}^{7}clamp(8\cdot I(x,y)-k,0,1)t_k \\ s.t.T_8=\sum_{i=0}^{7}t_i, T_k=\sum_{i=0}^{k-1}t_i \end{gathered} \tag{12}$$

where the $clamp(\ )$ is used to limit the input value to the range (0, 1).

In addition, histogram equalization is utilized to enhance the contrast of the image, which makes the image gray values approximately uniformly distributed by a nonlinear transformation. The basic principle is to find a suitable monotonic nonlinear mapping f to map the original image I_A to the target image I_B. To obtain the appropriate f, can be calculated as

$$p_B=f(p_A)=(L/A_0)\int_0^{p_A}H_A(p)dp \tag{13}$$

$$p_B=f(p_A)=(L/A_0)\sum_{k=0}^{p_A}H_A(k) \tag{14}$$



(a) TT100K (b) STSD (c) DFG

Fig. 4. Distribution of instance sizes and categories for the three traffic datasets. In order from left to right: TT100K, STSD and DFG.

where A_0 is the number of pixels, L=256 means the gray level depth, $H_A(\ )$ is the of the histogram distribution of original image, p denotes the size of the pixel values.

From the above method, the generalization of the histogram equalization filter can be obtained as written in (15), where the $HE(\ )$ represents the histogram equalization and α represents the parameter output by the Micro CNN.

$$I_{HE}(x,y)=\alpha\cdot HE(I(x,y))+(1-\alpha)\cdot I(x,y) \qquad (15)$$

Laplace sharpening is also used to highlight image details, and the generalization of the Laplace filter is written in (16).

$$I_L(x,y)=\begin{cases} I(x,y)+\lambda(I(x,y)-\nabla^2 I(x,y)), & \nabla^2 I(x,y)<0 \\ I(x,y)+\lambda(I(x,y)+\nabla^2 I(x,y)), & \nabla^2 I(x,y)>0 \end{cases} \qquad (16)$$

where ∇^2 represents the Laplace operator and the positive scale parameter λ is generated by Micro CNN, the sharpening degree of the filter can be adjusted by λ.

V. Experiments & Analysis

In this section, we provide a detailed and comprehensive evaluation of the performance of VATSD from multiple aspects. Firstly, we conduct a comparative analysis between VATSD and state-of-the-art (SOTA) methods in terms of overall performance, including detection accuracy and computational complexity. Secondly, we individually assess the reliability and efficiency of VATSD in addressing the three challenges proposed in this study: small target detection, complex working condition detection, and deployment on mobile devices. Lastly, through ablation experiments, we evaluate the impact of each module on model performance and further verify the ability of VATSD to achieve balanced performance across the three challenges.

A. Datasets

Five datasets that are currently more popular in the field of TSD were selected to evaluate the proposed method in this paper: Tsinghua-Tencent 100K (TT100K) [39], DFG [17], Swedish Traffic Signs Dataset (STSD) [10], CCTSDB [15], and CURE-TSD [7].

TT100K: TT100K provides 100k images of Chinese roads and contains a total of 30k traffic sign examples. Its categories are 221 in total, covering almost all traffic sign categories that appear in traffic scenes. TT100K dataset has a large proportion of small and medium objects, which is a good measure of the performance of small-sized traffic sign detection.

DFG: The DFG dataset contains 200 traffic sign categories captured on Slovenian roads, covering about 7,000 high-resolution images, each of which contains at least one instance of a traffic sign. The DFG is rich in traffic sign categories and has a large proportion of large objects.

STSD: STSD contains sequences from highways and others recorded from more than 350km of Swedish roads, and more than 20k images with 20% labeled. We select images containing traffic signs in the above three datasets that are labeled as training data. The light changes in STSD are more obvious, while its lower resolution can simulate the color distortion problems of actual traffic signs under five working conditions.

CCTSDB: The open source CCTSDB has 15723 images of Chinese roads, including different roads (urban, highway, and street) under different traffic scenario working conditions [40, 41]. However, the number of categories of CCTSDB is only 3, which is not applicable to evaluate the TSD. This dataset is used to evaluated IENet in this paper.

CURE-TSD: The CURE-TSD dataset released by George Institute of Technology contains 2 million traffic sign images based on real-world and simulator data. This dataset will increase the diversity of the dataset by adding a variety of complex working scenarios such as rain, snow, blur, and haze to the traffic sign images.

The first three datasets are utilized to evaluate the overall performance of VATSD, with their data distributions depicted in Fig. 4. These datasets are widely employed for the evaluation of traffic sign detectors due to their diverse categories and uniformly distributed sizes. While CCTSDB and CURE-TSD exhibit a lower number of traffic sign categories, they encompass diverse traffic scenarios, thereby facilitating the evaluation of detection capability in complex environments.

TABLE II
COMPARISON EXPERIMENTS WITH SOTA ON TT100K, STSD AND DFG

Datasets	Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
TT100K	RetinaNet	45.7	67.2	49.6	26.2	59.4	80.5
	QueryDet	57.1	77.1	66.3	37.9	69.8	79.6
	YOLOv5s	55.5	74.8	63.5	35.1	65.1	79.8
	OneNet	50.8	72.9	58.4	41.5	59.9	71.0
	FCOS	51.1	69.1	59.1	27.6	65.8	80.4
	YOLOX	60.6	80.3	71.7	40.7	66.9	75.7
	VATSD	**62.0**	**82.8**	**72.7**	**46.2**	**71.3**	**81.4**
	Faster R-CNN	59.4	78.5	70.4	42.7	70.2	77.4
	CenterNet	63.2	81.4	72.8	44.0	72.0	83.9
	Cascade R-CNN	**66.7**	**85.7**	**77.2**	45.6	**73.1**	**86.5**
	Sparse R-CNN	64.4	82.0	71.8	42.2	72.7	82.5
	VATSD	62.0	82.8	72.7	**46.2**	71.3	81.4
STSD	RetinaNet	51.7	78.9	53.8	43.8	69.6	80.4
	QueryDet	57.5	83.8	68.1	52.6	69.3	79.4
	YOLOv5s	58.6	83.5	70.4	50.4	67.8	81.2
	OneNet	60.1	86.3	73.3	56.8	66.9	78.3
	FCOS	53.6	76.2	62.0	39.5	75.8	86.8
	YOLOX	55.6	79.7	65.3	46.2	80.1	81.9
	VATSD	**65.9**	**89.6**	**76.6**	**59.2**	**81.9**	**91.0**
	Faster R-CNN	61.9	87.8	74.5	53.3	72.9	86.4
	CenterNet	65.2	**90.9**	77.0	54.2	76.3	89.5
	Cascade R-CNN	**68.4**	90.6	**78.4**	57.0	**84.0**	**91.3**
	Sparse R-CNN	63.7	88.3	75.6	53.9	82.5	88.1
	VATSD	65.9	89.6	76.6	**59.2**	81.9	91.0
DFG	RetinaNet	69.3	76.7	74.2	27.8	68.7	84.1
	QueryDet	72.5	79.4	77.0	25.3	62.3	85.8
	YOLOv5s	71.7	77.9	76.1	28.5	67.8	86.4
	OneNet	70.3	78.9	76.4	28.1	67.5	86.5
	FCOS	70.2	75.9	74.5	25.0	70.4	85.2
	YOLOX	73.1	82.5	88.1	31.6	70.5	87.6
	VATSD	**79.0**	**89.4**	**88.2**	**32.4**	**72.4**	**88.1**
	Faster R-CNN	77.8	86.0	86.1	27.9	70.8	87.2
	CenterNet	82.9	90.8	88.5	30.1	73.2	89.0
	Cascade R-CNN	**86.3**	**92.5**	**90.2**	30.8	**75.6**	**90.4**
	Sparse R-CNN	75.3	83.1	81.7	23.2	67.9	85.3
	VATSD	79.0	89.4	88.2	**32.4**	72.4	88.1

B. Experimental Details

Environment settings: Distributed experiments are used in that the training and testing of the model are carried out in two different hardware environments while ensuring that the training and testing environments of the comparison methods are uniform. The former training environment is as follows: Linux4.15.0-142-generic Ubuntu 18.04, with Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GH, 8×32GB DDR4 and 8×TITAN Xp, 12GB video memory, the batch size is set to 32. The test environment is as follows: Linux 5,13,0-40-generic Ubuntu 20.04.1, with Intel(R) Core (TM) i7-10700 CPU @ 2.90GHz, 2×16GB DDR4 memory and 1×GeForce RTX 3080 which with 10GB video memory.

Metrics: The evaluation metrics used in this section are as follows:

$$Precision(P) = \frac{TP}{TP+FP}$$

$$Accuracy(Acc.) = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Recall(R) = \frac{TP}{TP+FN}$$

$$AP = \int_0^1 p(r)dr$$

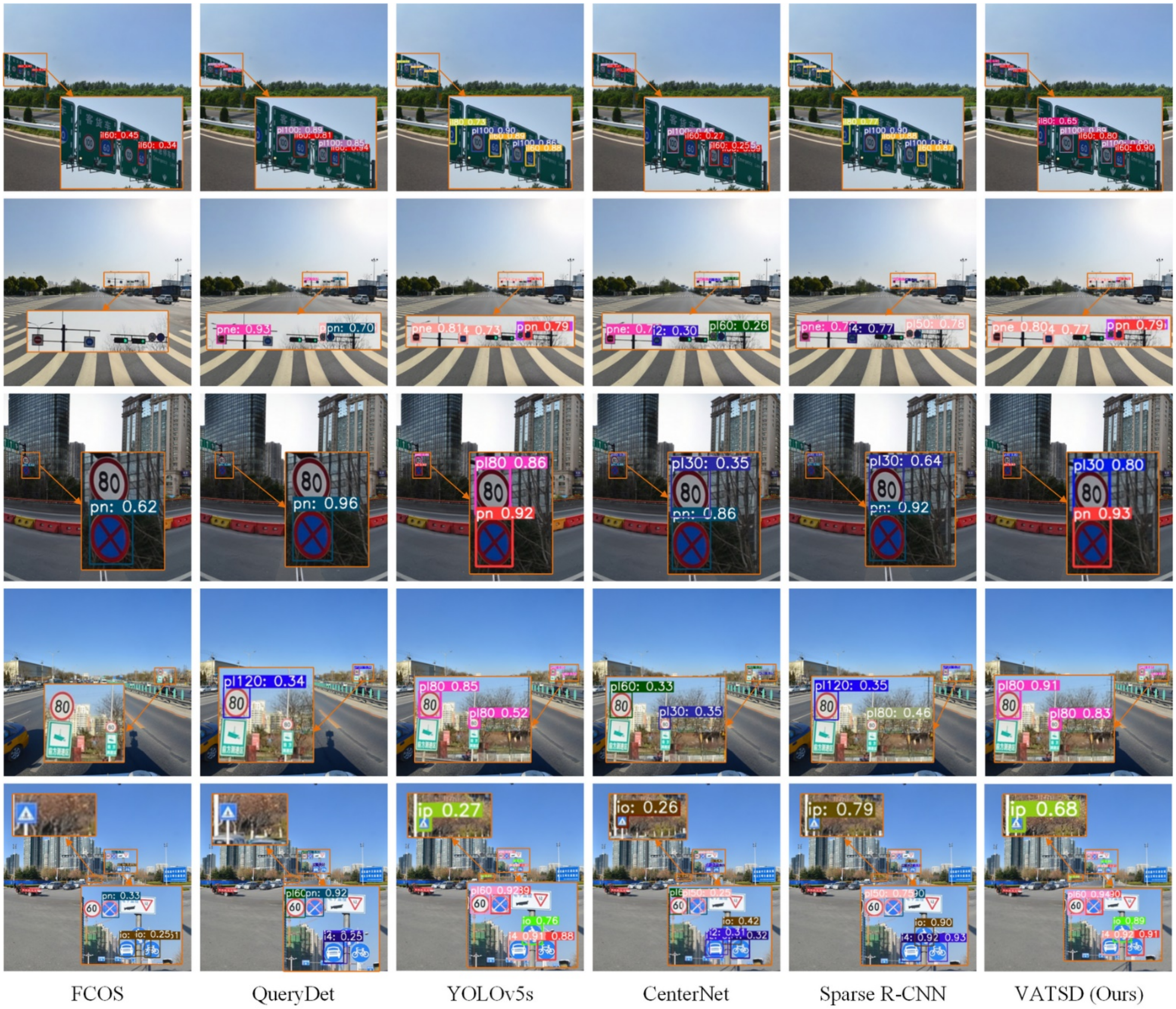


Fig. 5. Qualitative experimental results of the VATSD and SOTA methods on the TT100K dataset.

AP_{50}: AP at IoU=0.50

AP_{75}: AP at IoU=0.75

AP_S: AP at IoU=0.50:0.05:0.95 for small objects: area < 32^2

AP_M: AP at IoU=0.50:0.05:0.95 for medium objects: 32^2 <area< 96^2

AP_L: AP at IoU=0.50:0.05:0.95 for large objects: area > 96^2

$$FPS = \frac{1}{N}\sum_{n=1}^{N} f_n$$

TP (True Positive) denotes the number of samples predicted to be positive and actually positive, *TN* (True Negative) represents the number of samples predicted to be negative but actually negative. *FP* (False Positive) represents the number of samples predicted to be positive but actually negative, *FN* (False Negative) represents the number of samples predicted to be negative but actually positive. AP is computed through Precision-Recall Curve, $p(r)$ denotes the maximum precision at recall level.

Frames per second (FPS) is used to evaluate the inference speed of the model, indicating the number of images that can be processed per second. N denotes the number of images in the test set and f_n denotes the inference speed on the n-th image.

C. Performance Comparison with Other Methods

The proposed VATSD is compared with other state-of-the-art (SOTA) methods across three datasets: TT100K, STSD, and DFG, to evaluate its overall performance. The analysis will be conducted from both quantitative and qualitative perspectives.

Qualitative analysis: The detection accuracies of various methods are presented in Table II. RetinaNet [9], QueryDet [11], YOLOv5s [42], YOLOX [43], OneNet [22], and FCOS [24] are popular one-stage object detection networks in recent years. These networks employ lightweight network architectures, enabling real-time and efficient traffic sign detection, while also facilitating deployment on mobile devices. Faster R-CNN [5], CenterNet [3], Sparse R-CNN [44] and

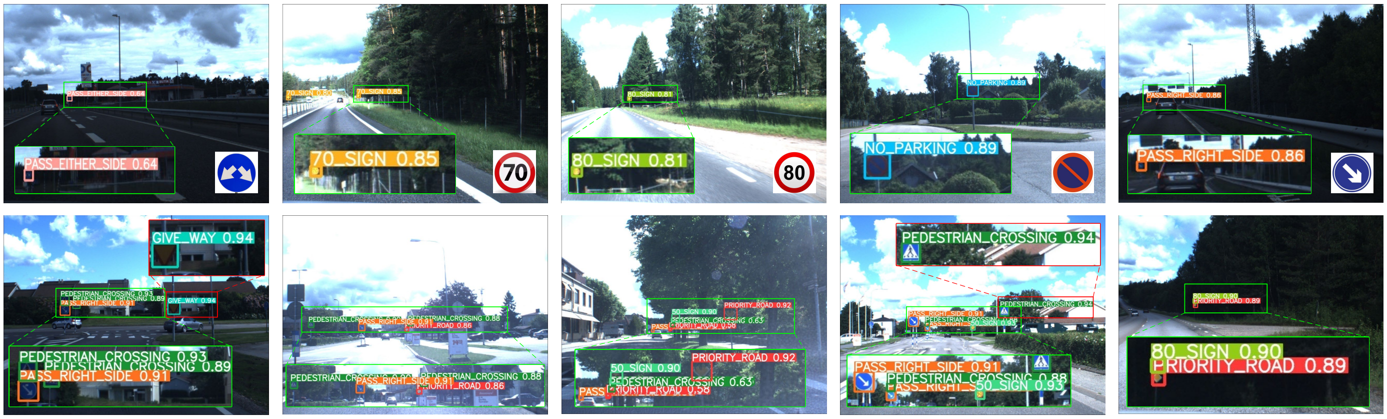


Fig. 6. Qualitative experimental results of the proposed method on the STSD dataset. Enlarged display of small-sized traffic signs in green and red boxes.

TABLE III
COMPARISON OF TIME COMPLEXITY, SPACE COMPLEXITY AND SPEED

Methods	T_C	S_C	Infer time	Acc.
QueryDet	37.6	38.3	0.093	57.5
FCOS	34.8	32.2	0.067	53.6
YOLOv5s	16.0	6.75	0.008	58.6
Cascade R-CNN	57.8	69.4	0.142	68.4
VATSD	16.6	7.86	0.010	65.9

*T_C: Time complexity (GB), S_C: Space complexity (MB), *Infer time (s)

TABLE IV
EVALUATION RESULTS ON TT100K-S

Method	P	R	F_1	mAP
RetinaNet	19.3	37.9	25.6	10.2
YOLOv5s	30.5	55.4	39.3	24.6
FCOS	18.4	41.0	25.4	13.5
Cascade R-CNN	32.7	63.3	43.1	29.8
VATSD	**46.1**	**67.8**	**54.9**	**34.8**

Cascade R-CNN [4] are two-stage object detection networks that achieve high detection accuracy through intricately designed network structures.

From Table II, VATSD outperforms the one-stage networks in terms of accuracy metrics on three datasets. Compared to the two-stage networks, VATSD also demonstrates strong competitiveness in terms of detection accuracy, with a minor difference compared to the well-performing Cascade R-CNN. Notably, VATSD excels in detecting small-sized objects and achieving 46.2%, 59.2%, and 32.4% on the TT100K, STSD, and DFG, respectively. The STSD dataset exhibits significant variations in brightness and low image resolution, making it difficult for the human eye to discern distant traffic signs. However, the proposed method shows improved detection accuracy compared to the TT100K dataset, indicating its adaptability to complex environments. Additionally, Table III showcases the evaluation results of each method concerning computational complexity, inference speed, and detection accuracy. From the Table III, VATSD exhibits a computational complexity that is closely ranked after YOLOv5s, with a minimal difference, and its inference speed is only slower by 2ms. Remarkably, VATSD achieves a detection accuracy that surpasses YOLOv5s by 7.3%. Cascade R-CNN notably outperforms other one-stage comparison networks in terms of detection accuracy, yet its margin compared to VATSD is merely 2.5%. While the time complexity and space complexity of VATSD are 16.6GB and 7.86MB respectively, which are far lower than Cascade R-CNN. These findings demonstrate that VATSD not only excels in detection accuracy but also demonstrates strong competitiveness in terms of computational complexity and inference speed.

Quantitative analysis: Fig. 5 illustrates the detection results of various methods on the TT100K dataset. From Fig. 5, our approach demonstrates more precise localization and recognition of traffic signs compared to other methods. Particularly, VATSD achieves accurate identification of distant traffic signs as well. In contrast, methods such as FCOS and QueryDet are prone to omission and false positive, as observed in rows three and four, where detection of the "pl80" category is concerned. Fig. 6 and 7 present the test samples of VATSD on the STSD and DFG datasets, respectively. The STSD dataset contains a large number of traffic signs and includes scenarios with varying exposure and dim lighting. In such cases, VATSD demonstrates high detection confidence (above 0.85) for signs like "NO_PARKING" and "PASS_RIGHT-SIDE," showcasing its reliability. The DFG dataset comprises 200 traffic categories, with certain signs bearing high visual similarity, leading to potential false positive detections. Through the DCSP module, VATSD effectively extracts precise target features and captures semantic information, enabling multi-class traffic sign detection.

D. Performance of Small-Sized Signs Detection

To further substantiate the high-precision capability of VATSD in detecting small-sized traffic signs, a specialized dataset called TT100K-S is constructed by leveraging the TT100K dataset. This created dataset only contains small-sized

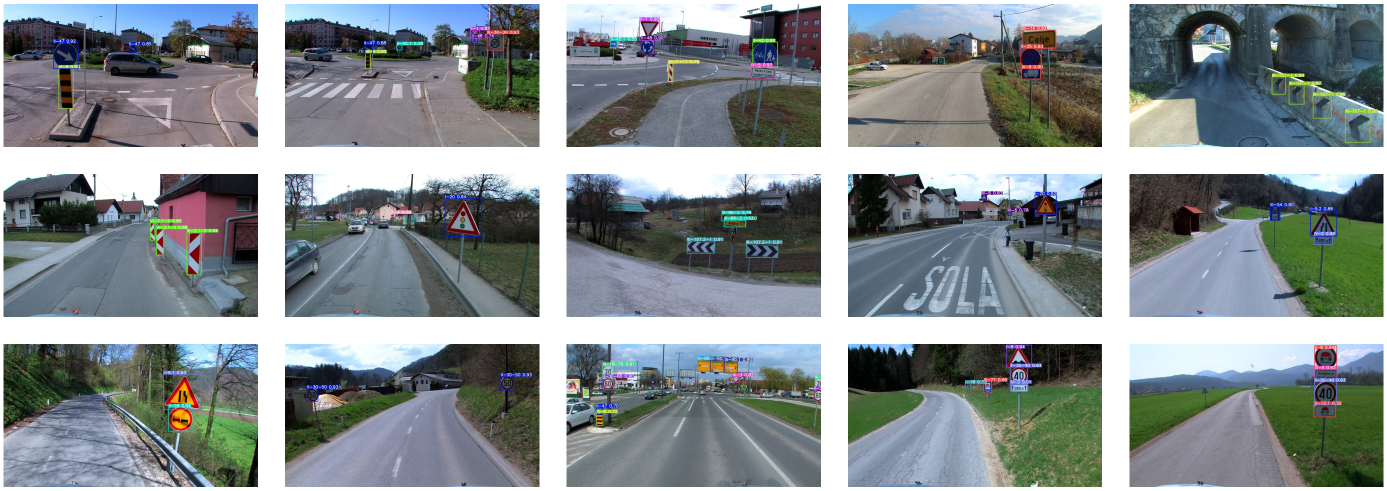


Fig. 7. Qualitative experimental results of the proposed method on the DFG dataset.

TABLE V

COMPARISON OF THE DETECTION ACCURACY OF EACH METHOD UNDER DIFFERENT WORKING CONDITIONS

Method	Sunny			Dimness			Fog/Haze		
	All	Speed limit	Stop	All	Speed limit	Stop	All	Speed limit	Stop
RetinaNet	30.2	42.8	45.6	11.8	24.6	26.6	16.4	28.3	32.2
YOLOv5s	34.4	47.7	51.8	16.3	27.0	30.2	19.5	31.1	35.8
FCOS	30.6	44.3	46.1	12.8	23.2	24.0	15.4	27.0	32.6
YOLOX	38.5	52.9	54.1	17.4	29.8	31.6	22.0	33.4	37.0
Sparse R-CNN	41.8	56.2	57.0	21.7	34.9	35.8	26.3	38.2	41.6
Cascade R-CNN	**47.5**	**55.7**	**60.5**	24.9	37.1	37.0	30.1	41.4	43.8
VATSD	44.2	53.4	56.8	**38.3**	**46.6**	**47.3**	**39.7**	**49.0**	**50.3**

traffic sign instances. By training VATSD on TT100K-S, we ensure that the proposed approach is not influenced by objects of different sizes, thereby accentuating its proficiency in detecting small-sized traffic signs. Since the dataset solely comprises small targets, we evaluated VATSD based on four key metrics: Precision, Recall, F_1, and mAP, as depicted in Table IV.

From Table IV, VATSD demonstrates superior performance across all metrics when compared to other methods. Here mAP is equivalent to AP_S because only small-sized traffic signs exist in the dataset. VATSD achieves an impressive mAP of 34.8%, surpassing the two-stage network Cascade R-CNN by a margin of 5%. Furthermore, VATSD exhibits considerably higher recall rate (67.8%) and F1 score (54.9%) than the one-stage network, indicating its effectiveness in mitigating both missed detections and false positives.

E. Performance in Complex Working Conditions

Firstly, we assess the detection performance of VATSD in both typical and challenging conditions and compare it with other methods, as presented in Table V. We consider sunny weather scenes as representative of typical conditions (favorable lighting and clear visibility for detection), while challenging conditions involve adverse weather such as rain and fog. To evaluate the generalization ability of VATSD, we employ a model trained on the DFG for testing.

Table V provides a comprehensive analysis, where the "All" represents the average detection accuracy across all classes for each method. Notably, we highlight the detection accuracy of the "Speed limit" and "Stop" classes, which are frequently occurring in the dataset, to emphasize the performance differences more prominently. From Table V, VATSD exhibits significant advantages in detection accuracy compared to other methods. In “Sunny” conditions, VATSD achieves detection accuracy second only to Cascada R-CNN, while in complex scenes such as "Dimness", the detection accuracy of VATSD has not decreased significantly compared with other methods. For example, Cascada R-CNN experiences a 22.6% decline in detection accuracy in dim conditions compared to sunny weather, VATSD only experiences a modest decline of 5.9%. Under the “Fog/Haze” conditions, VATSD achieves detection accuracies of 49% and 50.3% for the “Speed limit” and “Stop” categories, respectively, surpassing the two-stage network Sparse R-CNN by approximately 10%. These findings highlight the reliability and stability of VATSD in complex detection scenarios.

Secondly, to evaluate the detection performance of VATSD on small objects under complex conditions, TT100K dataset, which contains a significant number of small-sized objects, was

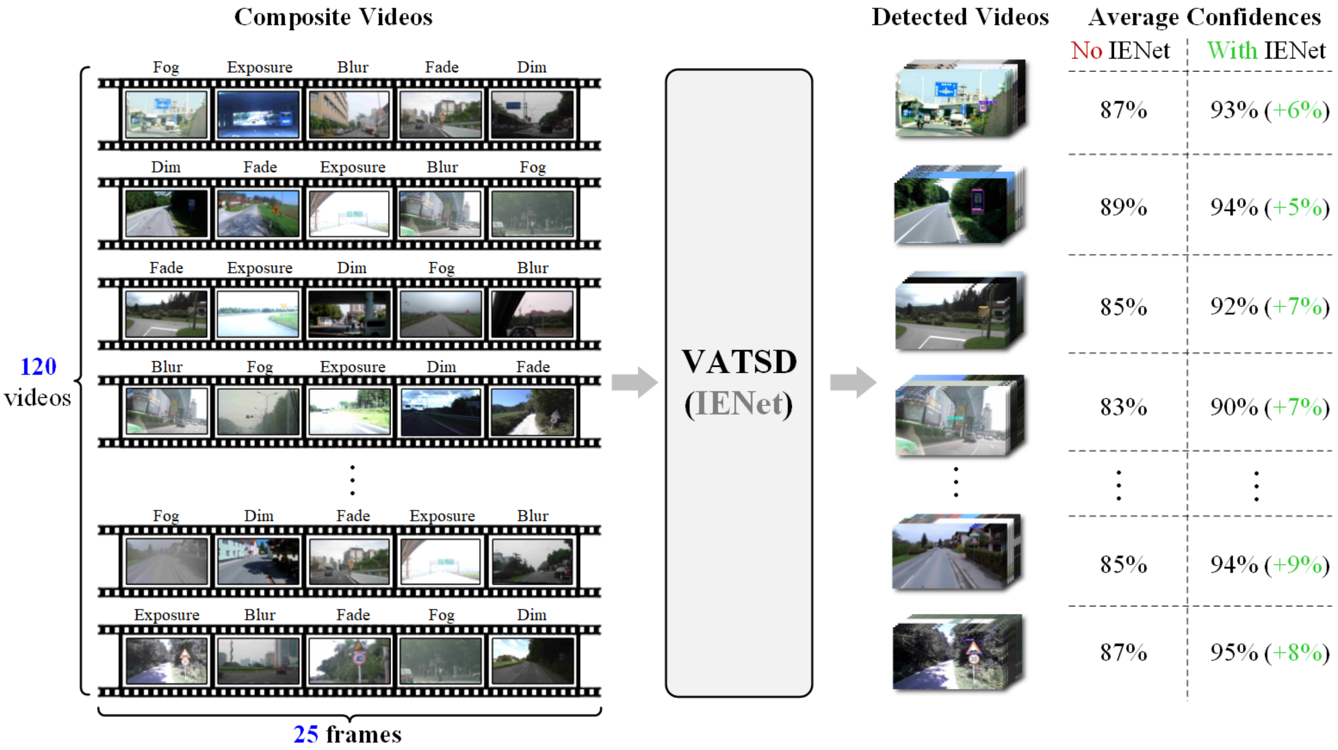


Fig. 8. Experimental results of IENet for adaptive image enhancement in dynamic scenes. Composite Videos denote the videos generated by the arrangement and combination of the five working scene images; Detected Videos denote the detection results of the videos by VATSD; Average Confidences denote the average confidence of VATSD with and without IENet.

TABLE VI
ROBUSTNESS ANALYSIS OF NOISE EFFECT

	No IENet		With IENet	
	AP	AP_S	AP	AP_S
Clean	61.2	45.4	62.0	46.2
Gaussian noise	56.3	32.7	61.5	43.2
Pepper noise	53.8	30.1	59.6	41.6
Speckle noise	50.0	26.9	59.1	37.9
Fog noise	53.4	37.4	61.6	40.4
High brightness	56.4	24.8	60.8	38.5
Low brightness	51.4	22.5	58.7	37.0

TABLE VII
VALIDATION OF IENET AND ADAPTIVE JOINT FILTERING STRATEGY

	IENet	AJF	Average Confidence	Time(ms)
Net.1			84%	/
Net.2	√		80%	7
Net.3	√	√	93%	3

subjected to various noise perturbations to simulate complex scenarios. Table VI presents the detection accuracy of VATSD with and without the incorporation of IENet under different noise augmentation techniques. IENet, employing adaptive filtering algorithms, effectively enhances image optimization and significantly improves the reliability of VATSD in complex conditions. Among the noise types, the maximum improvement in detection accuracy is observed under Speckle noise, where the AP increases by 9.1%. Notably, for small-sized objects, IENet demonstrates even more pronounced enhancements in detection accuracy, achieving a peak improvement of 14.5% in optimizing for Low Brightness conditions. The AP and AP_S of Gaussian noise reach 61.5% and 43.2%, respectively. Thesenumbers are only marginally reduced by 0.5% and 3% compared to the case without noise enhancement. Therefore, IENet greatly enhances the reliability and stability of VATSD in complex scenarios.

Thirdly, qualitative experiments were conducted on IENet using TT100K, STSD, DFG, and CCTSDB datasets. Fig. 8 illustrates how VATSD dynamically adapts and optimizes the images under five different complex working conditions to enhance detection accuracy. We randomly selected images of five working conditions as video frames, five images for each working condition, a total of 25 frames. Through rehearsal and combination, 120 video sequences can be formed for testing. This approach allows for a better demonstration of the flexibility and dynamic nature of the adaptive filtering strategy in IENet. Table VII serves as both a quantitative representation of Fig. 8 and an illustration of the significant role played by the adaptive joint filtering strategy in IENet. In the table, “Time (s)” denotes the processing time required for a single image, “AJF” refers to the adaptive joint filtering strategy. The adaptive joint

TABLE VIII
COMPARISON RESULTS DEPLOYED TO JETSON XAVIER NX

Method	FPS	Avg Power	General Conditions		Complex Conditions	
			Detected Rate	Eval.	Detected Rate	Eval.
Base (None)	-	489mW	-	-	-	-
QueryDet	16.1	11.6W	79.5	Normal	61.4	N/A
FCOS	17.3	12.2W	71.1	Normal	55.8	Normal
YOLOv5s	23.5	5.1W	82.7	Good	63.9	Normal
Cascade R-CNN	-	17.9W	-	N/A	-	N/A
VATSD	21.6	4.8W	86.3	Good	74.8	Good

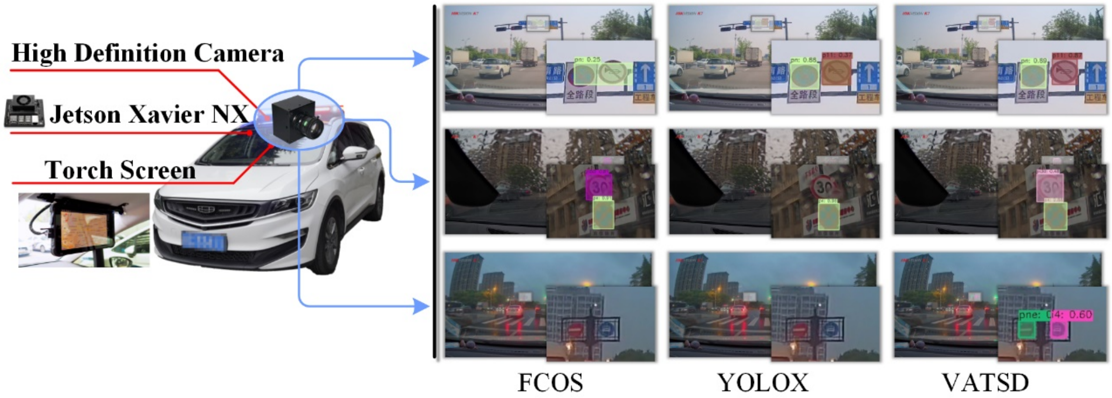


Fig. 9. Schematic diagram of the equipment on the vehicle side and on-road testing

TABLE IX
ABLATION EXPERIMENTS AMONG IENET AND DCSP

	+IENet	+DCSP	Acc.		
			TT100K	STSD	DFG
Net.1			57.3	58.8	71.4
Net.2		√	61.2	62.5	77.8
Net.3	√		58.6	63.7	73.1
Net.4	√	√	62.0	65.9	79.0

filtering strategy not only improves the speed of image processing but also enhances the average confidence level in the detection results, achieving a confidence level of 93%.

F. Performance on Mobile Devices Detection

Deploy the detection network on Jetson Xavier NX for actual road testing, as shown in Table VIII. "Avg Power" denotes the average power required when running the model on a mobile device, "Detected Rate" represents the ratio of correctly detected small objects (N_s) to the total number of samples (N_{all}). Due to the lack of ground truth in practical detection scenarios, objects with detection confidence exceeding 25% are generally considered as correct detections. All methods were tested under the same road scene conditions. "General Conditions" indicate well-illuminated and suitable detection scenarios, while "Complex Conditions" encompass scenarios with adverse weather conditions such as rain or fog that impact detection performance.

From Table VIII, it can be observed that the mobile device itself consumes 489mW of operational power, and VATSD requires 4.8W during runtime. This power consumption is lower than that of the lightweight network YOLOv5s, facilitating its deployment on diverse mobile devices. Furthermore, VATSD achieves a detection speed of 21.6FPS, second only to YOLOv5s, thereby meeting real-time detection requirements. VATSD exhibits a "Detected Rate" of 86.3% under general conditions and 74.8% under complex conditions, surpassing other one-stage networks. Conversely, the high-power consumption of Cascade R-CNN renders it unsuitable for normal operation.

Fig. 9 illustrates the detection samples of VATSD compared to other methods on mobile devices. We selected the well-performing one-stage networks YOLOX and FCOS for comparison. It can be observed that our method demonstrates excellent detection performance in sunny, rainy, and dim lighting conditions. In contrast, the other two methods exhibit instances of missed detections and positional deviations in rainy and dim lighting scenarios.

G. Ablation Experiments

This section validates the roles played by the proposed modules within VATSD. Table IX demonstrates the optimization effects of IENet and DCSP on the detection accuracy of VATSD. The feature extraction network in VATSD is composed of DCSP, while the networks labeled as Net.1 and Net.3 in the table utilize Darknet53 [45] as their feature extraction network without DCSP. The results on the three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TITS.2023.3309644, IEEE Transactions on Intelligent Transportation Systems

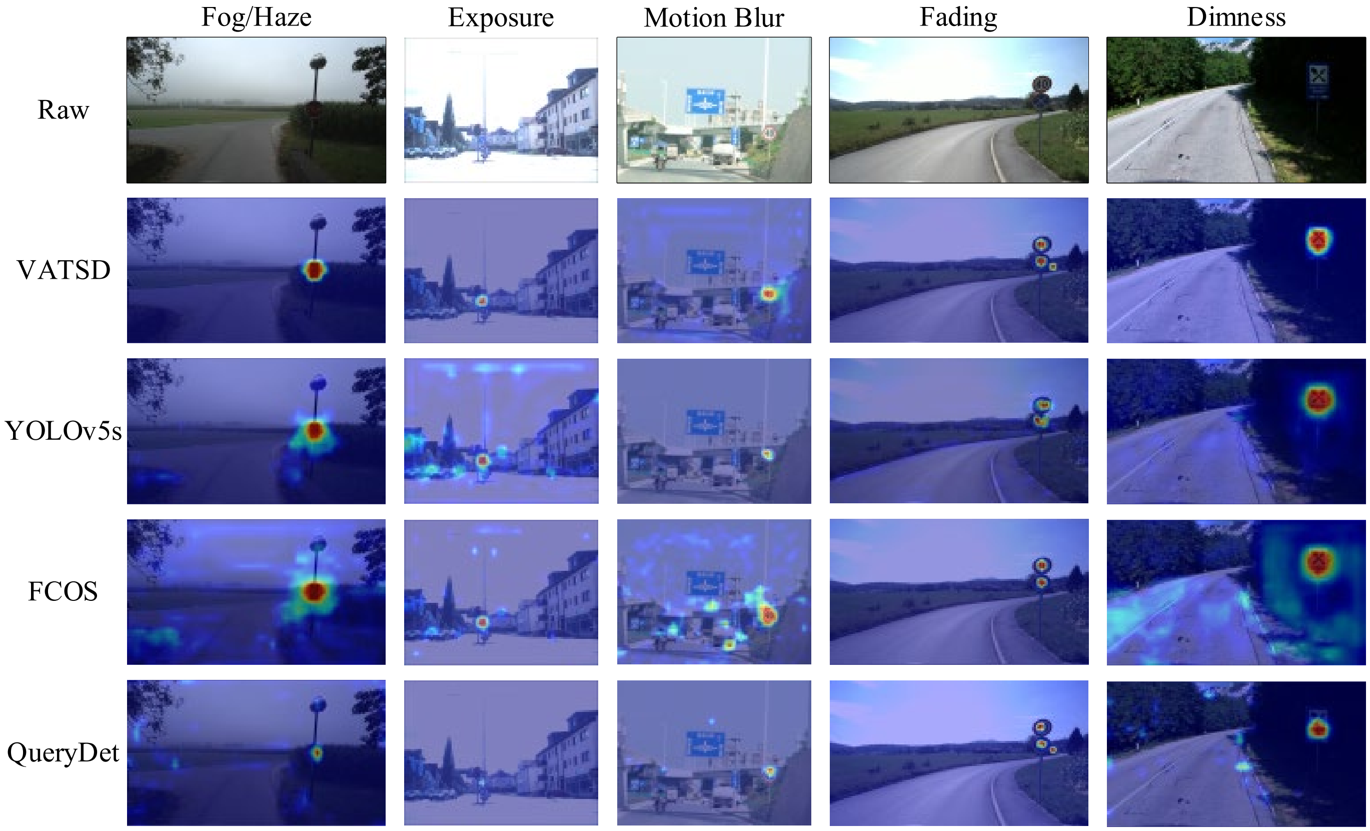


Fig. 10. Visualization of feature activations on the complex conditions. The red part indicates a high level of concern for the area

TABLE X
ABLATION EXPERIMENTS ON DIFFERENT CHANNEL SEPARATION RATIO

	x_0'	x_0''	x_0'''	Acc.		
				TT100K	STSD	DFG
Net.1	**2**	1	1	55.7	57.3	68.2
Net.2	1	**2**	1	59.3	61.0	73.9
Net.3	1	1	**2**	**61.2**	**62.5**	**77.8**

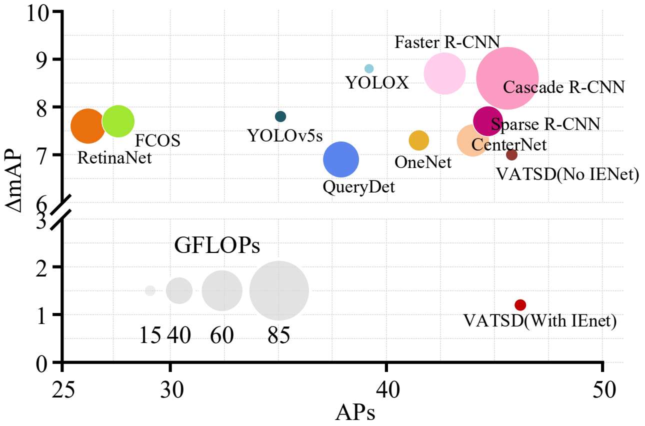


Fig. 11. Optimal balance between VATSD and SOTA approaches for small-scale traffic sign detection, adaptability to complex working conditions, and deployment on mobile terminals.

datasets indicate that both DCSP and IENet contribute to the improvement of the detection results to varying degrees. The STSD dataset contains complex traffic scenes, demonstrating the optimization effect of IENet in challenging scenarios, with an accuracy increase of 4.9%. Similarly, DCSP improves the feature extraction capabilities. It achieves an accuracy of 61.2% on the TT100K, which contains a significant number of small-sized objects, and exhibits a notable 6.4% improvement in accuracy on the DFG dataset, showcasing the largest improvement magnitude.

Meanwhile, we analyzed the channel division ratio of DCSP and obtained the optimal design method. Table X shows the detection performance under three channel divisions. Three ratios are set: 2:1:1, 1:2:1, 1:1:2. The experimental results show that the optimal detection results are obtained on the three data sets in the case of 1:1:2. We also conducted experiments on the extreme cases of channel splitting ratio. The extreme cases include setting a channel to 0 or setting the proportion of one channel much larger than the rest. In each of these extreme cases, the accuracy of the model deteriorates or improves only slightly.

Fig. 10 presents the feature visualization results of various methods. It can be observed that VATSD achieves precise localization and focus on the objects under five different complex conditions, while other methods exhibit certain errors. For instance, in the first column depicting a foggy scene, YOLOv5 and FCOS fail to accurately focus on the traffic signs.

In the second column, QueryDet exhibits insufficient feature extraction for small-sized traffic signs, leading to the attention being concentrated only on a subset of the traffic signs.

The main focus of this paper is to address three key challenges: high-precision recognition of small-sized objects, adaptability to complex working conditions, and deployment on mobile devices. Generally, improving detection performance often requires sacrificing certain computational and memory costs. Considering the balance among these factors, this paper proposes DCSP and IENet to optimize the model structure for lightweight design. Fig. 11 shows the model performance of VASTD and other methods facing the above three challenges. All detection methods are tested on the TT100K. The test set with randomly added noise is used to evaluate the adaptability to complex working conditions by measuring the variation of mAP before and after noise addition. AP_S is used to assess the accuracy of small object detection, while GFLOPs are used to evaluate the computational complexity of the models as a metric of the difficulty in deploying them on mobile devices.

From Fig. 11, VATSD effectively balances the aforementioned challenges, with a ΔmAP variation of no more than 2%, achieving high accuracy in small object detection while controlling model computation costs. Other methods, due to their failure to consider the complex working conditions faced by the detector, experience significant degradation in detection accuracy when confronted with noisy images. Although networks such as Cascade R-CNN and CenterNet demonstrate good performance in small object detection, their high computational costs hinder their deployment on mobile devices, making them unsuitable for practical needs.

VI. Conclusion

In this paper, VATSD for small traffic sign detection is proposed that not only trade-off the detection accuracy and detection speed of small traffic signs, but also enables efficient detection under a variety of complex working conditions. First, to improve the feature extraction capability of the detector the Backbone of the network is improved. A feature tight fusion structure is designed and the DCSP-based feature extraction network effectively improves the feature learning of small objects with negligible additional parameters. Secondly, the adaptive joint filtering strategy is proposed for complex working conditions, which efficiently realizes adaptive processing of different working conditions through the management of multiple filters. The proposed method is compared with other state-of-the-art methods on three traffic sign datasets. The results show that the proposed detector outperforms other methods in terms of small object detection accuracy under multi-conditions, with AP_S reaching 59.2% on the STSD dataset. Further, the VATSD is deployed on Jetson Xavier NX with a speed of 21.6 FPS, meeting the need for high accuracy and real-time traffic sign detection. In the future, we hope to design more comprehensive image enhancement networks for all the harsh environments faced by existing traffic sign detectors and to investigate efficient object detection algorithms using only the CPU.

References

[1] A. R. Javed, M. Usman, S. U. Rehman, M. U. Khan, and M. S. Haghighi, "Anomaly Detection in Automated Vehicles Using Multistage Attention-Based Convolutional Neural Network," *IEEE Trans. Intell. Transp. Syst.,* vol. 22, no. 7, pp. 4291-4300, Jul, 2021.

[2] W. D. Min, R. K. Liu, D. J. He, Q. Han, Q. T. Wei, and Q. Wang, "Traffic Sign Recognition Based on Semantic Scene Understanding and Structural Traffic Sign Location," *IEEE Trans. Intell. Transp. Syst.,* vol. 23, no. 9, pp. 15794-15807, Sept. 2022.

[3] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic Two-stage Detection," 2021, *arXiv:2103.07461.*

[4] Z. Cai, and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 43, no. 5, pp. 1483-1498, 2021.

[5] L. Yang, J. Zhong, Y. Zhang, S. Bai, G. Li, Y. Yang, and J. Zhang, "An Improving Faster-RCNN With Multi-Attention ResNet for Small Target Detection in Intelligent Autonomous Transport With 6G," *IEEE Trans. Intell. Transp. Syst.*, pp. 1-9, 2022.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788.

[7] D. Temel, M. H. Chen, and G. AlRegib, "Traffic Sign Detection Under Challenging Conditions: A Deeper Look into Performance Variations and Spectral Characteristics," *IEEE Trans. Intell. Transp. Syst.,* vol. 21, no. 9, pp. 3663-3673, 2020.

[8] L. Shen, L. You, B. Peng, and C. Zhang, "Group multi-scale attention pyramid network for traffic sign detection," *Neurocomputing,* vol. 452, no. 10, pp. 1-14, 2021.

[9] X. Cheng, and J. Yu, "RetinaNet With Difference Channel Attention and Adaptively Spatial Feature Fusion for Steel Surface Defect Detection," *IEEE Trans. Instrum. Meas.,* vol. 70, pp. 1-11, 2021.

[10] F. Larsson, and M. Felsberg, "Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition," in *Image Analysis*, Berlin, Heidelberg, 2011, pp. 238-249.

[11] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13658-13667.

[12] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10778-10787.

[13] Y. Y. Zhu, C. Q. Zhang, D. Y. Zhou, X. G. Wang, X. Bai, and W. Y. Liu, "Traffic Sign Detection and Recognition Using Fully Convolutional Network Guided Proposals," *Neurocomputing,* vol. 214, pp. 758-766, Nov 19, 2016.

[14] J. Chen, K. Jia, W. Chen, Z. Lv, and R. Zhang, "A Real-time and High-precision Method for Small Traffic-signs Recognition," *Neural Computing and Applications,* vol. 34, no. 3, pp. 2233-2245, 2022/02/01, 2022.

[15] J. Zhang, M. Huang, X. Jin, and X. Li, "A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2," *Algorithms,* vol. 10, no. 4, 2017.

[16] J. Yu, X. J. Ye, and Q. Tu, "Traffic Sign Detection and Recognition in Multiimages Using a Fusion Model With YOLO and VGG Network," *IEEE Trans. Intell. Transp. Syst.*, May 2, 2022.

[17] D. Tabernik, and D. Skočaj, "Deep Learning for Large-Scale Traffic-Sign Detection and Recognition," *IEEE Trans. Intell. Transp. Syst.,* vol. 21, no. 4, pp. 1427-1440, April. 2020.

[18] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1571-1580.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017, *arXiv:1704.04861.*

[20] X. Zhang, X. Y. Zhou, M. X. Lin, and R. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848-6856.

[21] Y. Zhu, and W. Q. Yan, "Traffic Sign Recognition Based on Deep Learning," *Multimedia Tools and Applications,* vol. 81, no. 13, pp. 17779-17791, 2022/05/01, 2022.

[22] P. Sun, Y. Jiang, E. Xie, W. Shao, Z. Yuan, C. Wang, and P. Luo, "What Makes for End-to-End Object Detection?," 2020, *arXiv:2012.05780.*

[23] A. C. Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, Zeng Yifu, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, xylieong, "ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations," 2022.

[24] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A Simple and Strong Anchor-Free Object Detector," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 44, no. 4, pp. 1922-1933, 2022.

[25] S. Zheng, Y. Wu, S. Jiang, C. Lu, and G. Gupta, "Deblur-YOLO: Real-Time Object Detection with Efficient Blind Motion Deblurring," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1-8.

[26] J. X. Liang, J. W. Wang, Y. H. Quan, T. Y. Chen, J. Y. Liu, H. B. Ling, and Y. Xu, "Recurrent Exposure Generation for Low-Light Face Detection," *IEEE Trans. Multimedia,* vol. 24, pp. 1609-1621, 2022.

[27] X. Yuan, X. L. Hao, H. J. Chen, and X. Y. Wei, "Robust Traffic Sign Recognition Based on Color Global and Local Oriented Edge Magnitude Patterns," *IEEE Trans. Intell. Transp. Syst.,* vol. 15, no. 4, pp. 1466-1477, Aug, 2014.

[28] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic Convolution: Attention over Convolution Kernels," 2019, *arXiv:1912.03458.*

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, *arXiv:1512.03385.*

[30] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261-2269.

[31] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "Scaled-YOLOv4: Scaling Cross Stage Partial Network," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13024-13033.

[32] C. Qi, I. Nyalala, and K. Chen, "Detecting the Early Flowering Stage of Tea Chrysanthemum Using the F-YOLO Model," *Agronomy,* vol. 11, no. 5, 2021.

[33] G. Wang, H. Ding, B. Li, R. Nie, and Y. Zhao, "Trident-YOLO: Improving the Precision and Speed of Mobile Device Object Detection," *IET Image Processing,* vol. 16, no. 1, pp. 145-157, 2022/01/01, 2022.

[34] S. J. Thorpe, "Image Processing by the Human Visual System," in *Advances in Computer Graphics*, Berlin, Heidelberg, 1991, pp. 309-341.

[35] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss Functions for Image Restoration With Neural Networks," *IEEE Transactions on Computational Imaging,* vol. 3, no. 1, pp. 47-57, 2017.

[36] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 33, no. 12, pp. 2341-2353, 2011.

[37] M. Ju, C. Ding, W. Ren, Y. Yang, D. Zhang, and Y. J. Guo, "IDE: Image Dehazing and Exposure Using an Enhanced Atmospheric Scattering Model," *IEEE Trans. Image Processing,* vol. 30, pp. 2180-2192, 2021.

[38] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, "Exposure: A White-Box Photo Post-Processing Framework," *ACM Trans. Graph.,* vol. 37, no. 2, 2018.

[39] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-Sign Detection and Classification in the Wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2110-2118.

[40] J. M. Zhang, X. Zou, L. D. Kuang, J. Wang, R. S. Sherratt, and X. F. Yu, "CCTSDB 2021: A More Comprehensive Traffic Sign Detection Benchmark," *Human-Centric Computing and Information Sciences,* vol. 12, May 30, 2022.

[41] J. M. Zhang, Z. P. Xie, J. Sun, X. Zou, and J. Wang, "A Cascaded R-CNN With Multiscale Attention and Imbalanced Samples for Traffic Sign Detection," *IEEE Access,* vol. 8, pp. 29742-29754, 2020.

[42] J. Wang, Y. Chen, Z. Dong, and M. Gao, "Improved YOLOv5 Network for Real-time Multi-scale Traffic Sign Detection," *Neural Computing and Applications*, 2022/12/09, 2022.

[43] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding Yolo Series in 2021," 2021, *arXiv:2107.08430.*

[44] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14449-14458.

[45] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions," 2021, *arXiv:2112.08088.*

Junfan Wang received the B.E. degree in Electornic Information Science and Technology from Wenzhou University, Wenzhou, Zhejiang, in 2020. She is currently working toward the Ph.D. degree in Integrated Circuit Science and Engineering from the School of Electoonic Science and Technology at the Hangzhou Dianzi University. Her research interests cover artificial neural network, intelligent transportation system and vehicle-road synergy

Yi Chen (Student Member, IEEE) received the M.E degree in Electronic Information in 2023 from the School of Electronic Information, Hangzhou Dianzi University, China. He is currently studying for a Ph.D. degree from the School of at Electrical Engineering, Zhejiang University, China. His research interests cover Computer Vision and AI Generation Content.

Xiaoyue Ji (Student Member, IEEE) received the B.E. degree in electronics and information engineering in 2016 from the School of Electrical Engineering, Harbin Engineering University, China. Shi is currently working toward the Ph.D. degree in control throry and control engineering from the School of Electrical Engineering, Zhejiang University, China. Her research interest cover artificial neural network and autonomous driving.

Zhekang Dong (Senior Member, IEEE**)** received the B.E. and M.E. degrees in electronics and information engineering in 2012 and 2015, respectively, from Southwest University, Chongqing, China. He received the Ph.D. degree from the School of Electrical Engineering, Zhejiang University, China, in 2019. Currently, he is an associate professor in Hangzhou Dianzi University, Hangzhou, China. He is also a Research Assistant (Joint-Supervision) at The Hong Kong Polytechnic University. His research interests cover artificial neural network and intelligent transportation system

Mingyu Gao was born in 1963. He received the M.S. degree in power electronics from Zhejiang University, Hangzhou, China, in 1993, and the Ph.D. degree in information and communication engineering from the Wuhan University of Technology, Wuhan, China, in 2013.

In 2001, he joined Hangzhou Dianzi University, Hangzhou, China, where he is currently a Professer with the School of Electronic and Information. His research interests include electronics and vehicle electronics.

Chun Sing Lai (Senior Member, IEEE) received his B.Eng. with first class honors in electronic and electrical engineering from Brunel University London, United Kingdom, and his D.Phil. in engineering science from the University of Oxford, United Kingdom, in 2013 and 2019, respectively. He is currently a lecturer in the Department of Electronic and Electrical Engineering, Brunel University London. His current interests are in data analytics and intelligent transportation system