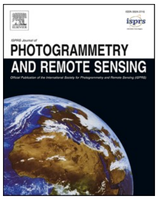
Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



SignHRNet: Street-level traffic signs recognition with an attentive semi-anchoring guided high-resolution network

Yongtao Yu[a,*], Tao Jiang[a], Yinyin Li[a], Haiyan Guan[b], Dilong Li[c], Lianghai Chen[a], Changhui Yu[a], Li Gao[a], Shangbing Gao[a], Jonathan Li[d]

[a] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, JS 223003, China
[b] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, JS 210044, China
[c] College of Computer Science and Technology, Huaqiao University, Xiamen, FJ 361021, China
[d] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L3G1, Canada

ARTICLE INFO

Keywords:
Traffic signs recognition
High-resolution network
Feature attention
Semi-anchoring
Feature alignment
Street-level image

ABSTRACT

Traffic signs are indispensable fixtures in modern transportation activities, which are installed along the roadside or over the pathway to provide important prompt messages. The timely emergences, complete visibilities, and definite signals of traffic signs matter significantly to direct the driving behaviors and ensure the convenience and security of the transportation activities. Moreover, the efficient and accurate recognition of traffic signs can also provide decisive evidences to a variety of intelligent transportation systems. Nevertheless, the unignorable presence issues of traffic signs caused by vision factors, such as small sizes, dimness, deformations, and occlusions, impact essentially on the high-quality recognition of traffic signs. This paper develops a novel attentive semi-anchoring guided high-resolution network, named SignHRNet, for street-level traffic signs recognition purpose. First, stacked with a high-resolution network boosted by a dual-attention module as the feature extractor, the SignHRNet can exploit informative channel features and task-oriented spatial features to generate multiscale strong feature semantics for instance-level predictions. Second, designed with a semi-anchoring guided detection strategy assisted by instance-aware feature alignment, the SignHRNet can achieve highly-efficient sign type categorization and highly-accurate sign location determination. The proposed SignHRNet is intensively evaluated on three large-size datasets towards traffic signs recognition. Quantitative verifications demonstrate a competitive accuracy with an average mAP, mAP50, and mAP75 of 72.85%, 96.48%, and 85.31%, respectively, in processing traffic signs of varying self-conditions under diverse scenarios. Ablative and comparative analyses also confirmed the practical reliability and performance superiority of the SignHRNet in traffic signs recognition applications.

1. Introduction

In modern road constructions, traffic signs are common types of infrastructures placed along the road corridors, which are leveraged to direct the transportation activities. Traffic signs function to transmit important signals to the road users and manage and control the driving behaviors with specifically designed colors, shapes, characters, and figures. With the accurate and timely information and guidance provided by the traffic signs, road users can succeed in rapidly approaching the destinations, as well as helping smooth the traffic flows and ensure the driving safeties. Furthermore, traffic signs can also provide the traffic management departments with necessary law enforcement evidences, such as violations of regulations. In summary, the visibility, readability, timeliness, and correctness of the information on the traffic signs matter essentially to guarantee the quality and security of the transportation activities. However, caused by anthropogenic or natural factors (e.g., traffic accidents, climates, and surrounding scenarios), traffic signs might undergo different-level damages and abrasions or be shielded by nearby objects, thereby resulting in the signal inaccuracy or unavailability. Hence, regularly performing traffic signs inspections and inventories to fix the anomalies and update the databases acts vitally to better serve the road-related applications, as well as providing the up-to-date auxiliary ingredients to the advanced driver assistance systems (Møgelmose et al., 2012; Yue et al., 2020). In addition, the detail-

[*] Corresponding author.
E-mail address: allennessy@hyit.edu.cn (Y. Yu).

https://doi.org/10.1016/j.isprsjprs.2022.08.011
Received 25 December 2021; Received in revised form 1 July 2022; Accepted 15 August 2022

accurate and inventory-complete traffic signs information can be also applied to the documentation of high-definition maps, the creation of virtual scenes, and the design of intelligent vehicles.

Considering the importance and wide use of traffic signs, intensive attentions have been drawn to carry out traffic signs measurements with increasingly advanced techniques. At the early stage, traffic signs measurements were primarily accomplished with the means of on-site manual operations by bridle-wise workers. Nevertheless, such solutions were inefficient and required a great amount of labors, especially when covering a large area of road networks. In recent decades, as the imaging sensors keep developing in resolutions and qualities, as well as the cost-effectiveness, image data have been positively used to assist in traffic signs measurement tasks. The rapid acquisition of traffic sign images along the road corridors can effectively reduce the consumption of manpower, more importantly, improving the work efficiency. Accordingly, numerous strategies have been put forward for the intelligent interpretations of road-scene images for high-performance traffic signs measurement purposes in the literature. Thereinto, traffic signs recognition is a hot topic and acts as an indispensable prerequisite to a broad range of applications. To be specific, traffic signs recognition comprises the tasks of traffic sign instances localization to identify the existences and traffic sign types determination to interpret the signals. As yet, a variety of algorithms have been designed for traffic signs recognition with excellent accuracies and efficiencies (Liu et al., 2019; Wali et al., 2019), some of which have even been applied to actual applications. However, on account of the diversities of imaging sensors, the illumination and weather conditions of the collected images, the situations of the traffic sign instances with regard to the sizes, shapes, types, deformations, viewpoints, and occlusions, and the complicated surrounding scenarios, the automation level and recognition accuracy of traffic signs still remain a certain gap in comparison with the human-level qualities. In other words, it is still remarkably challenging to achieve higher recognition accuracies and lower misidentification errors with totally automated processing schemes than those of manual recognitions by humans to handle traffic sign images of varying conditions. Therefore, the reliability and practicability should be further enhanced by investigating more advanced solutions in order to broaden the integrations and realize the values of traffic signs recognition modules in actual transportation-related applications.

To be specific, the motivation of this paper is to develop an effective model to achieve high traffic signs recognition accuracies, especially correctly recognizing the small-size traffic signs and the traffic signs under complicated scenarios and weather conditions, and develop an efficient model to achieve rapid traffic signs recognition speeds. In this paper, considering the processing efficiency and effectiveness, we design a novel one-stage architecture for carrying out traffic signs recognition in street-level scenes. This architecture involves a multi-branch high-resolution backbone for multiscale, task-aware feature semantic exploitation and a detection head for sign category determination and sign location prediction. Specifically, channel and spatial feature attentions are simultaneously taken into account for promoting the feature encoding quality and a semi-anchoring guided detection strategy, assisted by a feature alignment scheme, is proposed for efficiently obtaining high-quality bounding boxes of traffic signs. Noteworthily, the recognition accuracy improvement of the small-size traffic signs benefits significantly from the feature extraction backbone for maintaining a high-resolution branch across the entire network and the feature attention mechanism for highlighting the feature saliencies. The contributions of this paper are mainly embodied in the following aspects. (1) A lightweight, efficient, and effective dual-attention module is formulated for boosting the feature semantic quality and robustness by attending to the significant, informative channel features and salient, task-oriented spatial features. (2) A semi-anchoring guided strategy functioned with a feature alignment component is proposed by leveraging an anchor-free scheme for efficient sign types prediction and an anchor-based scheme for accurate bounding boxes prediction. (3) A large-size street-level dataset with detailed annotations is built for serving traffic signs recognition tasks.

The structure of this paper is formulated as follows. An intensive literature review on traffic signs recognition methods and feature attention techniques is presented in Section 2. The detailed architecture of the traffic signs recognition model is explained in Section 3. The quantitative and qualitative assessments are discussed in Section 4. Eventually, the conclusions and comments are summarized in Section 5.

2. Related works

2.1. Handcrafted feature based solutions

Since traffic signs are usually designed with specific shapes, colors, and patterns, conventional solutions often utilized such prior knowledge to conduct traffic signs recognition. Intensity based thresholding approaches, shape based matching strategies, semantic based segmentation schemes, and handcrafted feature based learning models are intently taken into account. Timofte et al. (2014) integrated the color and shape priors to roughly yield a candidate set of potential traffic sign regions, which were elaborately selected and distinguished through an AdaBoost detector. By considering the global color characteristics and local edge patterns, Yuan et al. (2014) designed an effective traffic signs recognition model. Specifically, color, spatial, direction, and shape properties were reasonably combined to enhance the model robustness. Khan et al. (2011) investigated a couple of shape criteria to identify traffic signs from the pre-segmented candidates, which were preliminarily determined using color cues. Later on, the sign types were subtly recognized through a reference based matching framework. On the issue of complex street scenarios, Guo et al. (2020) exploited the possibilities of color primitives, position priors, and structure parameters to locate text-form traffic signs. The textual signals with both vertical and horizontal permutations were further extracted via text line tracking. Likewise, Greenhalgh and Mirmehdi (2015) suggested a region-constraint strategy to rapidly locate the candidate searching scopes. The textual signals on the signs were interpreted by color thresholding and spatial text line formulation. Aiming at alleviating the challenges of handling distorted or small-size traffic signs, Yazdan and Varshosaz (2021) proposed a stereo image processing technique. First, potential traffic sign regions were segmented from the right image based on color features and classified to obtain the initial detection results using geometrical properties. Then, optimal recognition output was obtained through key point matching and skeleton construction with the assistance of the left image. To identity traffic signs of specific types, Ruta et al. (2010) proposed a prototype matching method to measure the similarity between a couple of images. The similarity measurement principle was established based on the SimBoost and regression tree models. Boumediene et al. (2014) exploited the transferable belief functions to encode region semantics for the pre-generation of test areas. Then, a template matching technique was applied to verify the traffic signs of interest.

As another research route, a number of studies dedicated to the design of semantic models or target classifiers on the basis of handcrafted feature encodings. Lu et al. (2012) developed a graph embedding formulation for modeling the intra-category variations and inter-category distinctions of traffic signs. The representation quality was significantly enhanced cooperated with a sparse learning technique. Liu et al. (2016) designed a cascaded tree detector to rapidly recognize the complete traffic signs from high-contrast region proposals, while the occluded signs were further discovered using a sparse representation based classifier. As for text traffic signs, González et al. (2014) employed a bag of visual words (BoVWs) representation to depict the semantic statistics of image features. The identification was finalized based on a support vector machine (SVM) classifier. In Yang et al. (2016), histogram of oriented gradient (HOG) features were considered to build the SVM classifier. Specifically, the region proposals of traffic signs were pre-located through color probability map analysis. In contrast, Greenhalgh and Mirmehdi (2012) constructed a group of SVM classifiers serving different tasks and integrated them in a cascaded manner to gradually obtain the sign locations and type details. In addition, channel feature architectures (Møgelmose et al., 2015), decision reasoning systems (Meuter et al., 2011), incremental frameworks (Yuan et al., 2017), and tree classifiers (Liu et al., 2014; Zaklouta and Stanciulescu, 2012) were also investigated for traffic signs recognition tasks. Macroscopically speaking, the prior cues or handcrafted features based solutions required a great deal of vigor to manually design the semantic rules and formulate the feature descriptors. The relaxation degrees of the rules and the representation robustness of the descriptors impact unignorably on the recognition performances, especially on the cases of challenging and varying circumstances.

2.2. Deep feature based solutions

Profiting from the development of hardware performances, deep learning architectures, typified by the convolutional neural networks (CNNs), have recently demonstrated superior advantages in a wide range of vision tasks. Roughly speaking, the mainstream detection models typically concentrate on the one-stage and two-stage processing pipelines. To be specific, the two-stage solutions involve the pre-generation of redundant object region proposals, which are further processed to provide finer recognitions. On the contrary, the one-stage solutions merely rely on a single forward regressor to yield the target parameters. Hence, due to the remarkable achievements and extensive attentions, deep learning models have also been intensively exploited for traffic signs recognition purposes. Zhu et al. (2018) stacked a pair of fully convolutional network (FCN) structures as the traffic signs segmentor and texts detector to, respectively, locate the text signs and fetch the signal details. Sun et al. (2020) formulated a single-shot detector, named Dense-RefineDet, to solve the small-size traffic signs identification issue. In this model, an anchor-refinement module exported optimized anchors to an object detection module for the purpose of promoting the localization accuracy. For lightweighting the model parameters, Song et al. (2019) designed an efficient CNN architecture to accelerate the detection of traffic signs in large contents. The efficiency promotion benefitted from the convolution pruning and substitution operations. Shen et al. (2021) integrated a feature attention mechanism into a pyramidal architecture to extract informative feature encodings at multiple scales, which contributed crucially to the detection of varying-size traffic signs. Faced with the phenomena of small sizes and occlusions, Liu et al. (2021) focused on the abstraction of scale-sensitive and context-aware feature representations with the assistance of an attentive pyramidal formulation and an adaptive context fusion module. To achieve traffic signs boundary delineations, Lee and Kim (2018) adopted the single shot multibox detector (SSD) to obtain the initial pose of a traffic sign, followed by a template transform to estimate the boundary corners. You et al. (2020) optimized the SSD architecture by applying small-size kernels to speed up the processing efficiency. The detection output was further enhanced based on spectral analysis and appearance transformation. Differently, Jin et al. (2020) improved the SSD model by fusing multilevel features and emphasizing salient feature channels, while Xie and Weng (2019) exploited multiscale and depthwise convolutions to boost the model robustness. As a representative family of one-stage detectors, you only look once (YOLO) architectures were paid close attentions towards traffic signs recognition. Gao et al. (2020) embedded the Gaussian mixture model into the YOLOv3 to serve for anchor clustering, thereby resulting in high-quality regression parameters. Tai et al. (2020) connected a spatial pyramidal pooling module into the YOLOv3 for boosting the multiscale feature saliencies, which was beneficial for the identifications of varying-size traffic signs. In addition, YOLOv4 and its variants (Avramović et al., 2020; Dewi et al., 2021; Wang et al., 2021) were also considered and applied to fulfil the real-time traffic signs recognition demands.

Aiming at enhancing the localization accuracies and recall rates, the two-stage models generated a quantity of redundant region proposals to cover the potential traffic sign instances, which were specifically tested to filter out the fake ones. Luo et al. (2018) adopted a spectral channel-based region segmentation approach to extract rough region proposals that possibly contained traffic sign instances. After proposal refinement, traffic signs were distinguished from the background using a multi-task CNN model. Kamal et al. (2020) formulated a segmentation network with the combination of the SegNet and U-Net, named SegU-Net, to extract the region candidates. These region candidates were post-processed through a CNN classifier to examine the presences of traffic signs. Wei et al. (2020) treated the recognition issue as a detection and classification task and cascaded a pair of CNNs to realize the goals separately. Specifically, the detector achieved the target localizations based on center point evaluations. In Zhang et al. (2020), region proposals were determined using a region proposal network (RPN) under multiple feature spaces. The recognition results were refined via a classification and regression CNN. Li and Wang (2019) combined the faster R-CNN with the Hough transform technique to elaborately delineate the sign contours. In this framework, Hough line and circle transforms cooperated to refine the bounding boxes on the assumption of shape priors. Shao et al. (2019) suggested an improved version of the faster R-CNN by optimizing the RPN with wavelet and extremal analyses. For handling small-size traffic signs, cross-stage features were reasonably fused for saliency preservation. Alternatively, Cao et al. (2021) promoted the multiscale feature encodings with a high-resolution network (HRNet) backbone alongside with an attentive sample selection scheme. Zhou et al. (2021) investigated the capability of the attention mechanism in enhancing the feature representation quality. The attention module operated parallelly to take into account the multiscale feature subspaces. The entire network followed the Libra R-CNN architecture. In contrast, Wang et al. (2020) designed an inception-based attention module functioning to access and integrate multiscale contextual semantics. As a strategy for narrowing the searching ranges, location and size priors were employed to build a probability distribution model. To better handle oriented text traffic signs, Bagi et al. (2022) constructed a position-sensitive oriented RPN, which can produce rotated anchors for tightly enclosing arbitrarily-oriented traffic signs. For feature boosting, channel selection and attention were embedded together with separable convolution operations. Serna and Ruichek (2020) proposed a processing pipeline of detection, refinement, and classification. To be specific, the mask R-CNN was first leveraged to supervise traffic signs detection; then, location cues were applied to filter out false alarms; finally, the sign types were determined based on a CNN classifier. As for challenging image conditions, Ahmed et al. (2022) built an encoder-decoder structure to recover the details. The enhanced image was fed into a proposal localizer to crop the sign contents, which were directly input to a category classifier for sign types recognition. In addition, few-shot models (Zhou et al., 2020), capsule networks (He et al., 2021), semi-supervised learning architectures (Nartey et al., 2020), transfer learning techniques (Mannan et al., 2019), extreme learning formulations (Zeng et al., 2017), and multi-source data fusion strategies (Guan et al., 2020; Javanmardi et al., 2021) were also exploited for traffic signs recognition tasks.

Comparatively, the deep features based models usually show excellent robustness on different image sources and reliable performances under varying instance conditions and scene variations. This is the key factor that they have attracted good graces in a wide range of vision tasks. However, the constructions of these models generally require large numbers of annotated samples and a bulk of computation resources, which more or less restrict their applications to some extent in some special cases.

2.3. Attention mechanisms in CNNs

In order to further promote the feature representation quality to improve the prediction accuracies of the vision tasks, many attempts have been recently made to strengthen the contributions of the useful feature semantics (Guo et al., 2022). Roughly speaking, existing techniques generally focus on the recalibrations of the channel feature semantics to highlight the task-specific channels and the recalibrations of the spatial feature semantics to emphasize the task-specific regions. Hu et al. (2020) developed a squeeze-and-excitation (SE) block to adaptively recalibrate the channel feature semantics. The SE block determined the channel-wise significances by modelling the interdependencies among the channels. As a modification, Zhang et al. (2021) proposed a pyramid squeeze attention (PSA) module for channel feature promotion under different scales. Specifically, the PSA module took the multiscale channel features as the input and accomplished feature recalibration based on the SE block. Differently, Qin et al. (2021) presented a frequency channel attention module to exploit channel features under different frequencies. In this module, the feature semantics from different frequencies were concatenated and comprehensively considered to determine the channel feature saliencies. Hou et al. (2021) designed a coordinate attention block by embedding positional attributes into channel attentions. The coordinate attention block investigated the feature significances along the horizontal and vertical directions, respectively, which were finally combined to form the position-aware feature encodings. To integrate both local and global contents, Zhong et al. (2020) constructed a squeeze-and-attention (SA) module. Different from the SE block, the SA module employed a non-fully-squeezed scheme to parse the local feature details. Jaderberg et al. (2015) pioneered a spatial transformer network (STN) architecture to recalibrate feature semantics in the spatial domain. The STN contained three main components, including a localization net, a grid generator, and an image sampler, to determine an affine-transformation-invariant feature representation of the semantic target. Almahairi et al. (2016) developed a dynamic capacity network (DCN) formulation to adaptively assign the feature significances to different image portions. The selection was determined based on a gradient-based attention mechanism. To improve localization accuracy, Mayo et al. (2021) proposed a spatial embedding principle by using attention mechanisms. Through reinforcement learning, the built attention probability map was applied to infer the spatial information. Ulutan et al. (2020) employed a spatial graph network structure to exploit the relative spatial and structural correlations between the semantic objects. The spatial attention was achieved by learning the spatial interaction patterns between the object pairs. Aiming at realizing relative saliency encodings to highlight the foreground regions, Fang et al. (2021) suggested a position-preserved attention strategy. The attention module comprised a position embedding stage for enriching the feature semantics with positional properties and a feature interaction stage for making use of the mutual features between object proposals.

As a hybrid type of feature attention mechanisms, multiple feature attention schemes have been combined in some researches. Woo et al. (2018) designed a convolutional block attention module (CBAM) to simultaneously attend to the semantic-related channel and spatial features. The two subparts were cascaded to sequentially recalibrate the channel and spatial feature semantics. As an alternative, Fu et al. (2019) developed a dual-attention (DA) module by paralleling a position attention unit and a channel attention unit. These two units served, respectively, to emphasize the task-aware spatial and channel feature semantics, which were eventually fused to enhance the feature representation quality. Differently, Zhao and Wu (2019) applied the channel and spatial attention mechanisms, respectively, to different levels of features to conduct feature recalibrations separately. The attentive multilevel feature semantics were finally combined for directing predictions. Guo et al. (2021) proposed a separable self-attention module (SSA), which sequentially modeled the spatial and temporal correlations. By embedding the spatial contexts into the temporal modelling, the localization accuracy was significantly promoted. Chen et al. (2020) combined the feature attention with the confidence attention to optimize the model robustness. Specifically, the confidence attention scheme was applied to formulate the loss function for supervising the model training. To model long-range dependencies, Wiles et al. (2021) suggested a co-attention module to match feature semantics with precise spatial location evidences. The attention information was computed by comparing the similarities between feature pairs. Similarly, Feng et al. (2021) also employed a co-attention strategy by considering the channel and spatial feature saliencies. In addition, residual attention (Zhu and Wu, 2021), depth-sensitive attention (Sun et al., 2021), dynamic visual attention (Wang et al., 2019b), domain attention (Wang et al., 2019a), and vision transformers (Chen et al., 2021; Angles et al., 2021) were also investigated to perform feature attentions.

3. Methodology

Aiming at realizing competitive processing efficiency and advantageous recognition accuracy, the proposed model is designed as a one-stage semi-anchoring guided architecture. Fig. 1 shows the proposed attentive semi-anchoring guided high-resolution network (SignHRNet) developed for the traffic signs recognition task. The SignHRNet comprises two main components: a feature extraction backbone and a detection head, which are directly connected for feature encoding and target detection, respectively. The feature extraction backbone employs an HRNet formulation to exploit feature semantics in different subspaces. The detection head adopts a semi-anchoring guided strategy to narrow the searching spaces by focusing on salient regions. Noteworthily, a novel feature attention module is integrated into the HRNet backbone for boosting the multiscale feature semantics and a bounding box refinement process is mounted on the detection head for promoting the localization accuracy.

3.1. HRNet feature extraction backbone

A distinctive and powerful attribute of the HRNet architecture (Sun et al., 2019) is that it parallels a set of convolution branches to simultaneously explore different feature subspaces with different spatial resolutions towards high-level feature semantics extraction. Specifically, cross-branch information exchange is repeatedly carried out to provide each feature subspace with a global perspective of the feature contents under different spatial resolutions, thereby effectively enhancing the representation quality of the output features in each branch. Thus, in our architecture, we formulate the feature extraction backbone as the HRNet structure on the purpose of providing high-quality feature semantics.

As illustrated by the left part in Fig. 1, the HRNet backbone is structurized by four parallel branches, which, respectively, function to extract feature encodings with multiple stages in different subspaces under certain spatial resolutions. Specifically, a stage in a branch involves the convolutional layers between two successive cross-branch feature exchange processes. From top to bottom, the spatial resolutions of these branches are gradually reduced. Within each branch, the sizes and spatial resolutions of the feature maps are totally selfsame, which favors to guarantee the localization accuracy. Generally, the high-resolution branches make for the saliency preservations of the small-size instances, whereas the low-resolution branches are beneficial to the characterizations of the large-size instances. In our architecture, the top branch (Branch 1) maintains the same spatial resolution as the input image, and the other lower branches are successively zoomed out with a scaling coefficient of 0.5. Specifically, by maintaining the high-resolution branch of Branch 1 across the entire network, it is significantly beneficial to the feature characterization and recognition of the small-size traffic signs. To conduct cross-branch feature exchange, at the end of each stage, the higher-resolution branches are downscaled to the desired size of the target branch through strided convolutions and the lower-resolution branches are upscaled to the expected size of the target branch through deconvolution convolutions. These scale-adjusted features are concatenated with the copied features from the target branch

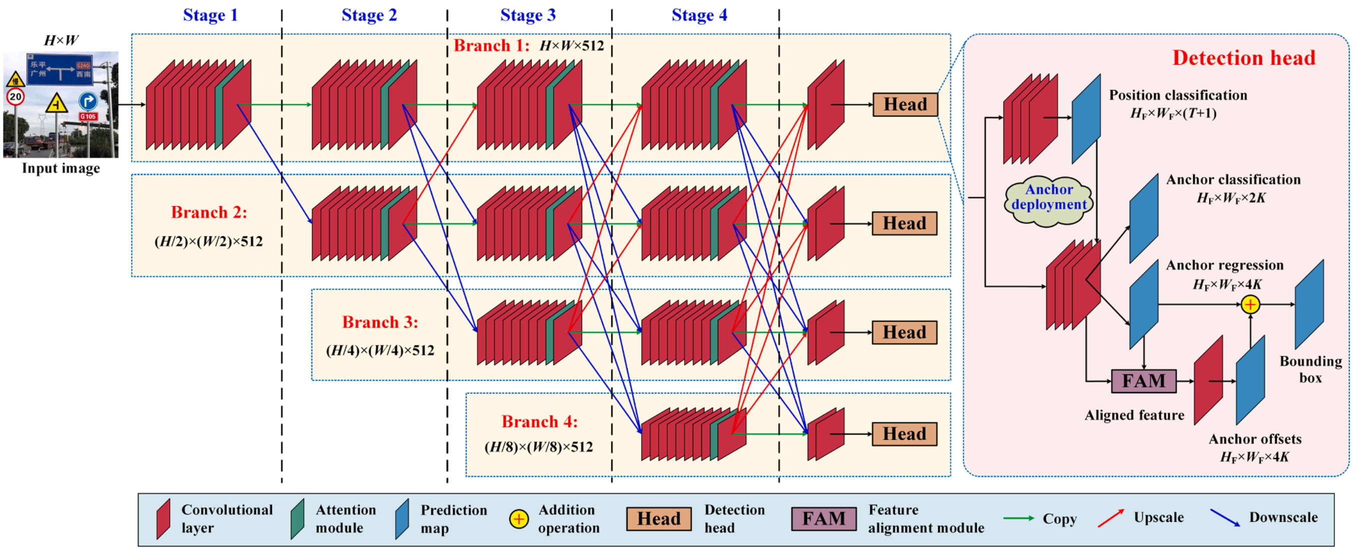


Fig. 1. Architecture of the proposed traffic signs recognition model.

and fused by a 1 × 1 convolution to form the semantic-promoted features in the next stage of the target branch. Eventually, the generated multiscale feature maps from all the four branches are leveraged as the semantic evidences for traffic signs recognition based on the detection head.

3.2. *Feature attention module*

As a matter of fact, convolution operations employ a sliding window pattern to interpret the feature semantics within the receptive field by using a certain size of convolution kernel. However, there is a fly in the ointment by performing only the pure convolutions. On the one hand, the distinctiveness and significance of different feature channels, particularly those closely associated with the foreground, are not specially considered and focused on, which impedes the extraction of high-quality, recognizable feature representations due to the equal contributions of the feature channels. On the other hand, the saliency and emphasis of the spatial positions, especially the ones covering the

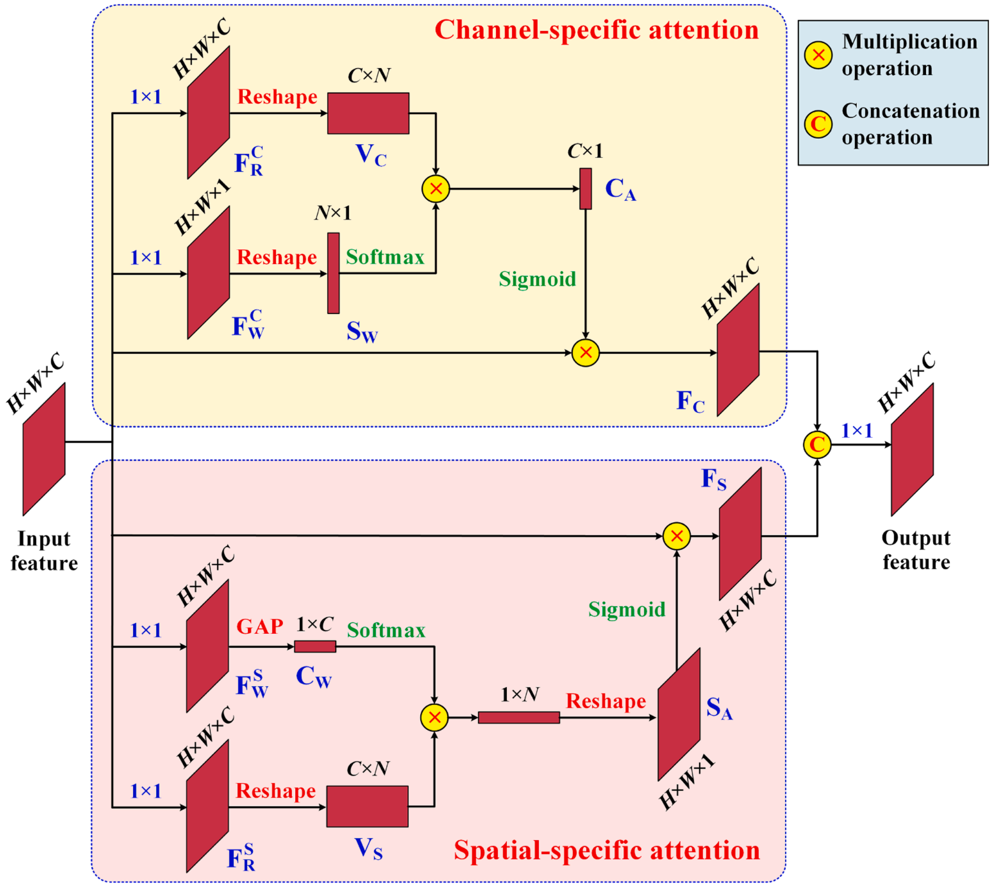


Fig. 2. Architecture of the proposed dual-attention module.

foreground, are not excellently considered and concentrated on, which goes against the extraction of robust, task-oriented feature representations due to the equal contributions of the spatial positions. Thus, recent studies have endeavored to design effective feature attention mechanisms aiming at enhancing the convolution performances by explicitly recalibrating the contributions of different feature semantics. For the same purpose, to further boost the quality and robustness of the generated multiscale features from the HRNet backbone, we build a novel dual-attention module and embed it into each branch of the HRNet backbone to supervise feature extraction. As shown by Fig. 2, the dual-attention module involves two components: a channel-specific attention unit and a spatial-specific attention unit, which, respectively, serve for recalibrating the channel and spatial feature contributions to highlight more important feature semantics. These two units operate parallelly on the input feature map and combine their outputs to form the quality-boosted feature map that preserves the identical channel number and spatial size. The design of the dual-attention module is inspired by the dual-attention module proposed by Fu et al. (2019), which parallels two branches to recalibrate spatial and channel feature semantics, respectively. However, the novelty and difference are that the proposed dual-attention module employs a more effective feature attention mechanism for obtaining higher-quality feature semantics and designs a more efficient lightweight architecture with less convolution parameters and less complex matrix operations. In addition, the dual-attention module contributes significantly to emphasize the feature saliencies of the small-size traffic signs, thereby improving the recognition accuracy of these small-size traffic signs.

As illustrated by the top path of the dual-attention module, the channel-specific attention unit achieves channel feature recalibration by upgrading the contributions of the informative channels. Concretely, given the input feature map with the dimension of $H \times W \times C$, where H, W, and C denote the height, width, and number of channels, respectively, first, two 1×1 convolutions are performed separately on the input feature map to produce a feature map$\mathrm{F}_{\mathrm{R}}^{\mathrm{C}} \in \mathrm{R}^{H\times W\times C}$ and a feature map$\mathrm{F}_{\mathrm{W}}^{\mathrm{C}} \in \mathrm{R}^{H\times W\times 1}$. Here, $\mathrm{F}_{\mathrm{R}}^{\mathrm{C}}$ can be treated as a feature response map, where each position encodes the task-aware feature responses related to the same position on the input feature map. While, $\mathrm{F}_{\mathrm{W}}^{\mathrm{C}}$ can be regarded as a spatial weighting map that reflects the importance of the feature responses at each position. To facilitate channel feature informativeness determination, $\mathrm{F}_{\mathrm{R}}^{\mathrm{C}}$ is reshaped to a feature matrix $\mathrm{V}_{\mathrm{C}} \in \mathrm{R}^{C\times N}$ and $\mathrm{F}_{\mathrm{W}}^{\mathrm{C}}$ is reshaped to a feature vector $\mathrm{S}_{\mathrm{W}} \in \mathrm{R}^{N\times 1}$, where $N = H \times W$. Next, V_{C} is multiplied with S_{W} based on normal matrix multiplication operations, resulting in a channel attention vector $\mathrm{C}_{\mathrm{A}} \in \mathrm{R}^{C\times 1}$. Specifically, the softmax function is applied to S_{W} before carrying out matrix multiplication to normalize the weighting factors. Each entry of C_{A} is computed by weighted aggregating the feature responses in each channel of $\mathrm{F}_{\mathrm{R}}^{\mathrm{C}}$ with a comprehensive evaluation of their significances, therefore it reflects the informativeness of the corresponding channel of the input feature map. To well portray the channel feature informativeness on the same baseline, the sigmoid function is applied to C_{A} to limit its encodings within one. Finally, by multiplying each of the attention factors in C_{A} to all the positions in the corresponding channel of the input feature map, we attain a quality-boosted feature map $\mathrm{F}_{\mathrm{C}} \in \mathrm{R}^{H\times W\times C}$, which explicitly attends to the important and informative channel feature semantics.

As illustrated by the bottom path of the dual-attention module, the spatial-specific attention unit achieves spatial feature recalibration by highlighting the saliencies of the foreground regions. Likewise, given the input feature map, two 1×1 convolutions are operated solely to generate a feature map$\mathrm{F}_{\mathrm{R}}^{\mathrm{S}} \in \mathrm{R}^{H\times W\times C}$ and a feature map$\mathrm{F}_{\mathrm{W}}^{\mathrm{S}} \in \mathrm{R}^{H\times W\times C}$. Notice that, $\mathrm{F}_{\mathrm{W}}^{\mathrm{S}}$ is post-processed by a global average pooling (GAP) operation to squeeze it into a feature vector $\mathrm{C}_{\mathrm{W}} \in \mathrm{R}^{1\times C}$. Here, $\mathrm{F}_{\mathrm{R}}^{\mathrm{S}}$ can be treated as a feature response map that encodes the task-sensitive feature responses at each position of the input feature map. While, C_{W} can be regarded as a channel weighting map that reflects the feature relevance of the feature responses in each channel. To facilitate spatial feature saliency determination, $\mathrm{F}_{\mathrm{R}}^{\mathrm{S}}$ is reshaped to a feature matrix $\mathrm{V}_{\mathrm{S}} \in \mathrm{R}^{C\times N}$, where $N = H \times W$. Then, V_{S} is multiplied by C_{W} through normal matrix multiplication operations to intently investigate the inter-channel dependencies. Specifically, the softmax function is applied to C_{W} before performing matrix multiplication to normalize the weighting coefficients. Next, after reshaping the product matrix along a row manner, we attain a spatial attention map $\mathrm{S}_{\mathrm{A}} \in \mathrm{R}^{H\times W\times 1}$, where each entry is computed by comprehensively taking into account the dependencies of the feature responses in all the channels, therefore, it reflects the saliency of the corresponding position on the input feature map. Similarly, to well portray the spatial feature saliency on the same baseline, the sigmoid function is applied to S_{A} to limit its encodings within one. Finally, by multiplying each of the attention factors in S_{A} to the corresponding position in each channel of the input feature map, we attain a quality-boosted feature map $\mathrm{F}_{\mathrm{S}} \in \mathrm{R}^{H\times W\times C}$, which explicitly attends to the salient and task-oriented spatial feature semantics.

As shown in Fig. 2, the recalibrated feature maps F_{C} and F_{S} generated by these two parallel attention units are directly concatenated and organically fused via a 1×1 convolution to form the final quality-boosted feature map that concurrently attends to both the channel and spatial useful feature semantics. As illustrated by Fig. 1, the dual-attention module is integrated in each branch of the HRNet backbone to boost the feature semantics used in the cross-branch feature exchange procedure. Concretely, the dual-attention module is mounted at the end of each stage in each branch of the HRNet backbone, i.e. at the point before performing cross-branch feature exchange in each stage.

3.3. Detection head

Considering the processing efficiency of the traffic signs recognition model, the detection head is devised to be a one-stage formulation. In regard to the one-stage solutions, the anchor-free architectures demonstrate advantageous efficiencies, but they might suffer from the slight inaccuracy of the regressed target parameters. On the contrary, the anchor-based architectures demonstrate competitive accuracies, but they undergo the substantial computation overhead in regressing the large quantities of anchors, even a considerable portion of useless anchors are distributed in the background regions. Hence, targeting at combining the strengths of the anchor-free structures in processing efficiency and the anchor-based structures in regression accuracy, we propose a semi-anchoring guided architecture as the detection head. Concretely speaking, first, we employ an anchor-free strategy to directly verify the positions on the feature map to identify the region candidates (i.e., the foreground positions) where the traffic signs possibly reside. Then, we adopt an anchor-based strategy to deploy a set of predesigned anchors at only the foreground positions to perform geometric parameters regression. With such a design pattern, the total number of anchors deployed on the feature map is dramatically reduced. In summary, the superiorities of the semi-anchoring guided architecture are embodied in the following two aspects. First, by narrowing the searching scopes and processing lightweight anchors, the detection head can achieve compatible efficiency with the anchor-free architectures. Second, by regressing geometric parameters with high-quality anchors, the detection head can achieve compatible accuracy with the anchor-based architectures. Noteworthily, the semi-anchoring guided architecture is quite different from the RPN used in the faster R-CNN (Ren et al., 2017). Specifically, the RPN deploys dense anchors at each position of a feature map and sorts and selects the high-quality regressed anchors to generate a set of region proposals, which are further classified to conduct object detection. In contrast, the novelty and difference of the semi-anchoring guided architecture are that it deploys anchors only at the predicted foreground positions rather than the entire feature map and directly uses the anchors to regress the object bounding boxes for object detection without generating any region proposals, which significantly improves the processing efficiency while achieving promising recognition accuracy.

As illustrated by the right part in Fig. 1, the detection head is composed of two branches: an anchor-free position classification branch and an anchor-based bounding box regression branch. The position classification branch predicts a category map through a shallow convolutional subnetwork, where each position outputs a $(T + 1)$-dimensional softmax prediction associated with the T categories of traffic signs and the background, respectively. That is, the entry with the maximum softmax output figures out the category label of a position. Then, the positions confirmed to belong to the traffic sign regions constitute the foreground positions, which will be treated as the reference searching scopes and attentively considered to deploy anchors for supervising bounding boxes regression.

The bounding box regression branch convolves a shallow subnetwork to generate a multi-task feature map, which is used for bounding boxes prediction in an anchor-based manner. Concretely, as illustrated by Fig. 1, a set of K predesigned anchors with different sizes and aspect ratios are distributed only at the foreground positions, which are predicted by the position classification branch, on the multi-task feature map. Similar to that in the faster R-CNN (Ren et al., 2017), the predesigned anchors are manually determined based on the prior knowledge of the traffic signs and quantitatively evaluated through experiments. The design of the anchors should take into consideration the size and aspect ratio variations of the traffic signs in the images. By default, we use three sizes and three aspect ratios, resulting in $K = 9$ anchors at a position. Based on the multi-task feature map and the deployed lightweight anchors, the bounding box regression branch first produces two outputs for anchor classification and anchor regression, respectively. The anchor classification terminal functions to examine whether an anchor at a foreground position optimally encloses a traffic sign instance. Therefore, the anchor classification terminal outputs $2K$ prediction scores (each anchor involves two softmax outputs) at each position to deduce the objectness probability (i.e., optimally enclosing a traffic sign instance or not) for each anchor. As shown in Fig. 3(a), the bounding box of a traffic sign is parameterized by a quaternion representation (x, y, w, h), which signifies a horizontal box with a width w and a height h centered at position (x, y). Therefore, based on such a bounding box parameterization, the anchor regression terminal outputs $4K$ predictions at each position for representing the four regression parameters of each of the K anchors. Eventually, by combining the prediction results from the anchor classification and anchor regression terminals, the bounding boxes of the traffic sign instances can be attained. That is, if an anchor at a position is confirmed to optimally enclose a traffic sign instance by the anchor classification terminal, the corresponding regressed parameters at the same position by the anchor regression terminal form the bounding box of the traffic sign.

As a matter of fact, by applying the same convolution operations with the same receptive fields, the multi-task feature map used for bounding boxes prediction provides the same scale of feature semantics and the same size of feature contexts at all positions across the entire feature map, which weakly exploits the variations and distinctiveness of the traffic sign instances at different positions to provide instance-oriented optimal feature encodings. As a result, with the weakly focused feature encodings, the regressed bounding boxes might be slightly inaccurate on the cases of small-size traffic signs or the traffic signs under complex scenarios (Fig. 4(a)). Ideally, a large-size instance should access a large feature context to focus on a large receptive field for obtaining a complete feature representation, whereas a small-size instance should only involve a small feature scope to focus on a small receptive field for reducing the impacts from the background features. Thus, as shown in Fig. 1, aiming at further optimizing the predicted bounding boxes, we propose a feature alignment module (FAM) for recalibrating the multi-task feature map based on the primary bounding boxes to generate an enhanced instance-oriented aligned feature map that focuses tightly on instance regions. As illustrated by Fig. 5, the FAM takes the bounding box regression map from the anchor regression terminal and the multi-task feature map as the input and performs a 3 × 3 deformable convolution operation to generate the aligned feature map. Specifically, the offset field used in the deformable convolution is generated based on the bounding box regression map by applying a 1 × 1 convolution operation. Then, the aligned feature map is leveraged to predict the anchor offsets, which are used to adjust the primarily regressed bounding boxes. That is, the anchor offset output involves $4K$ predictions at each position corresponding to the four offset parameters of each of the bounding boxes. Finally, the predicted anchor offsets are directly added to the primarily regressed bounding boxes from the anchor regression terminal to generate the final set of optimized bounding boxes prediction results (Fig. 4(b)).

Noteworthily, to finalize traffic signs detection, the non-maximum suppression process is applied only to the optimized bounding boxes at the foreground positions to remove those redundant overlapping bounding boxes belonging to the same traffic sign. In this way, the total number of bounding boxes being processed is dramatically reduced. Then, the objectness scores predicted by the anchor classification terminal are leveraged to determine the traffic signs from the remaining bounding boxes at the foreground positions.

3.4. Loss functions

The construction and optimization of the SignHRNet should be supervised with well-annotated ground-truth label maps. Since there are four sets of outputs in the detection head: one for the position classification branch and three for the bounding box regression branch, four sets of ground-truth label maps should be coupled for directing the backpropagation process. However, the two sets of outputs in the

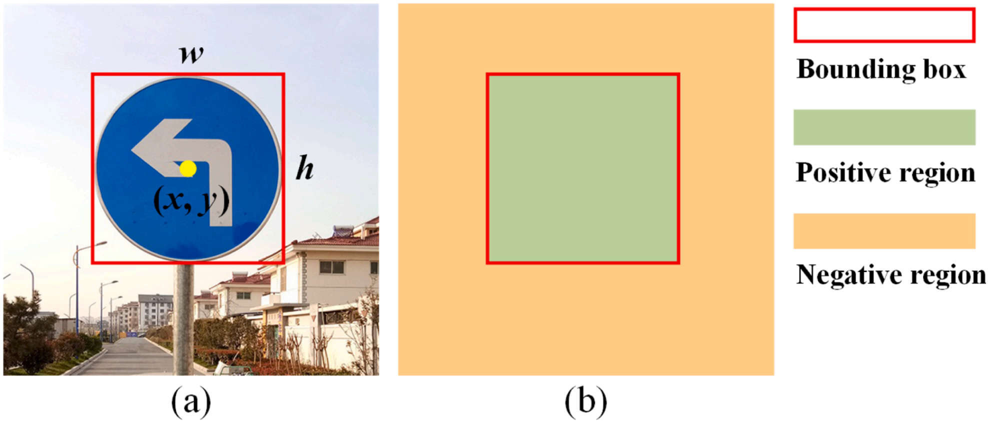


Fig. 3. Illustrations of (a) the parameterization of the bounding box and (b) the bounding box-based partition of the positive and negative regions.

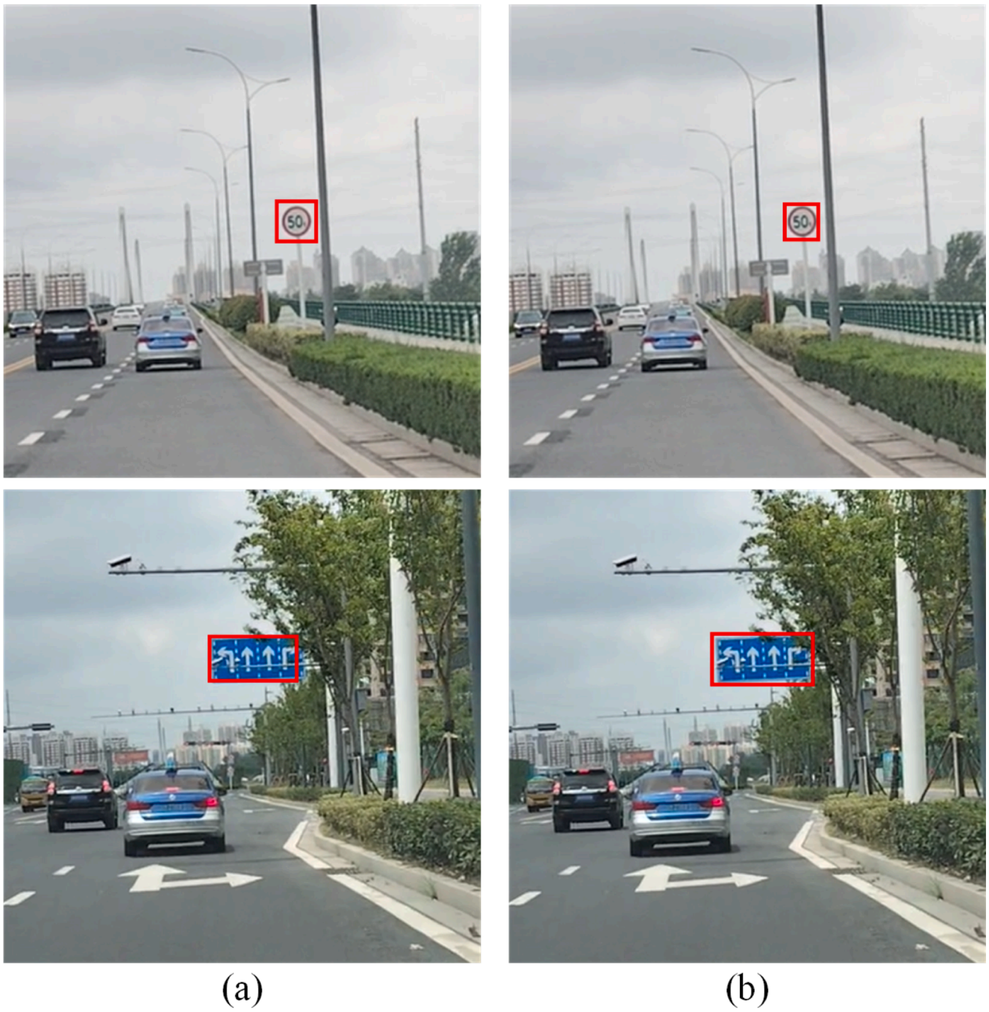


Fig. 4. (a) Primarily regressed bounding boxes and (b) optimized bounding boxes.

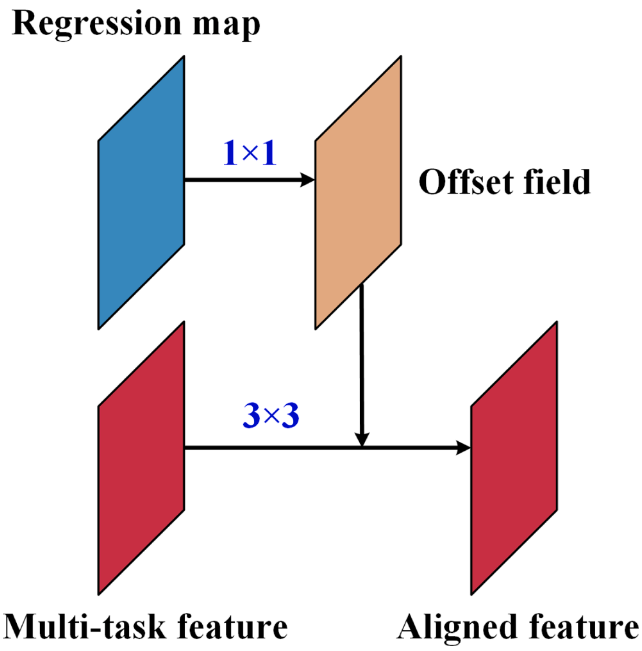


Fig. 5. Architecture of the feature alignment module.

bounding box regression branch pursue the same objective of bounding boxes regression, thus, they share the same set of ground-truth label map. For the position classification branch, a category label map that marks the positive positions associated with the traffic signs and the negative positions associated with the background is provided. As shown by Fig. 3(b), the regions enclosed by the bounding boxes are marked as the positive positions, whereas the external background regions are marked as the negative positions. Based on this category label map, the loss function used for optimizing the position classification branch is designed as the focal loss (Lin et al., 2020) as follows:

$$L_{pos} = \sum_i -(1-p_i)^2 \log p_i \tag{1}$$

where p_i denotes the softmax prediction corresponding to the ground-truth category entry marked in the category label map at position i.

For the bounding box regression branch, the ground-truth bounding box parameters are provided as the references. Specifically, the loss functions used for optimizing the anchor regression terminal and the optimal bounding box prediction terminal have the same formulation and are designed as the combination of two loss terms as follows:

$$L_{reg} = L_{smooth-L_1} + L_{GIoU} \tag{2}$$

where $L_{\text{smooth - }L_1}$ denotes the smooth-L_1 loss (Girshick, 2015) between the regressed parameters and the ground-truth parameters for serving for parameter-wise regression supervisions and L_{GIoU} denotes the generalized intersection over union (GIoU) loss (Rezatofighi et al., 2019) between the regressed bounding box and the ground-truth bounding box for serving for bounding box-wise regression supervisions. In addition, the loss function used for optimizing the anchor classification terminal is also designed as the focal loss (Lin et al., 2020) as follows:

$$L_{\text{cls}} = \sum_{a} -(1-p_a)^2 \log p_a \tag{3}$$

where p_a denotes the binary softmax prediction (i.e., positive anchor or negative anchor) at a foreground position corresponding to the ground-truth category entry of anchor a.

4. Results and discussions

4.1. Datasets

As benchmarks to verify the performance of the proposed Sign-HRNet, as well as providing comparative analyses with the other models used in this paper, we conducted intensive evaluations on two publicly released traffic signs recognition datasets and a large-size traffic signs recognition dataset built in this paper. The details of these three datasets are described in the following.

Tsinghua-Tencent 100K (TT-100K) (Zhu et al., 2016). This dataset involves 100,000 images, which were collected by using the Tencent Street View panoramas. The images exhibit with different weather conditions and illumination conditions. There are a total of about 30,000 traffic sign instances of different sizes and self-conditions from 45 categories. Specifically, 90 % of the images in this dataset are background images containing no traffic sign instances. The traffic signs in this dataset were annotated with horizontal bounding boxes and associated with the corresponding category labels. All the images have the same resolution of 2048 × 2048 pixels.

Challenging unreal and real environments for traffic sign detection (CURE-TSD) (Temel et al., 2019). This is a large-scale dataset containing about 1.72 million video frames with controlled synthetic challenging conditions. The video frames are of 1236 × 1628 pixels in resolution. The video frames are basically categorized into two types: real data and unreal data. The real data correspond to the frames collected from the real-world scenes, while the unreal data correspond to the synthesized frames generated in a virtual environment. To be specific, there are 49 real video sequences and 49 unreal video sequences, respectively, which are called challenge-free data without any post-processing. The 49 challenge-free real video sequences were further post-processed with 12 different kinds of effects and 5 different levels of challenges, thereby resulting in 2989 real video sequences. Furthermore, the 49 challenge-free unreal video sequences were post-processed with 11 different kinds of effects and 5 different levels of challenges, thereby generating 2744 unreal video sequences. The traffic signs in all the frames were annotated with horizontal bounding boxes as the localization references.

In-vehicle images for traffic sign recognition (IVI-TSR). This is a large-size dataset that was specifically built in this paper for traffic signs recognition tasks. This dataset includes 80,000 images, which were captured by an OPPO A93 mobile phone installed inside a Buick vehicle (Fig. 6(a)). The images were collected on the roads in Huaian, Jiangsu, on the roads in Nantong, Jiangsu, and on the highways between Huaian and Nantong. Each of the images has a size of 4000 × 3000 pixels and contains at least one traffic sign instance. The traffic signs cover 35 different categories with great variations in illumination and weather conditions, as well as under diverse surrounding scenarios and self-conditions. As shown in Fig. 6(b), a traffic sign is annotated with a horizontal bounding box that tightly encloses the traffic sign and tagged with a category label. The annotations of the images were manually accomplished by twelve undergraduate students from Huaiyin Institute of Technology by using the labelme tool (https://github.com/wkentaro/labelme). After annotations, double check was carried out by another five postgraduate students from Huaiyin Institute of Technology to correct the annotation errors by including the missing unannotated traffic sign instances and correcting the inaccurate bounding boxes or the wrong category labels of the traffic sign instances. Specifically, to provide a versatile benchmark dataset for different evaluation choices, the IVI-TSR dataset was not divided into the training, validation, and test sets. In summary, the main differences between the IVI-TSR dataset and the existing datasets lie in the following aspects. (1) The images were captured using a mobile phone sensor. (2) The images were collected in different places with different road environments, including urban and suburban roads, as well as highways. (3) The images covered different illumination and weather conditions, especially at dusk or early night and in rain and fog weathers. (4) The traffic signs exhibited varying self-conditions, especially occlusions caused by nearby objects. (5) The road scenes contained many advertising boards showing very similar appearances to the traffic signs.

For each dataset, 70 % of the data were randomly selected as the training set including 10 % of the data as the validation set, and the remaining 30 % of the data formed the test set for performance evaluation. This data partition scheme was applied to all the architectures and their repeatedly trained variants used in this paper. Specifically, the input sizes of the images to all the architectures used in this paper were configured as 1024 × 1024 pixels, 760 × 1000 pixels, and 1000 × 750 pixels, respectively, for the TT-100K, CURE-TSD, and IVI-TSR datasets by considering the GPU memory.

4.2. Parameter setting and model training

In all the experiments, a cloud computing platform was used for model construction and testing. This platform was equipped with a 128-GB memory, a 16-core CPU, and ten 16-GB GPUs. The proposed Sign-HRNet was trained using the Adam optimizer in an end-to-end way for

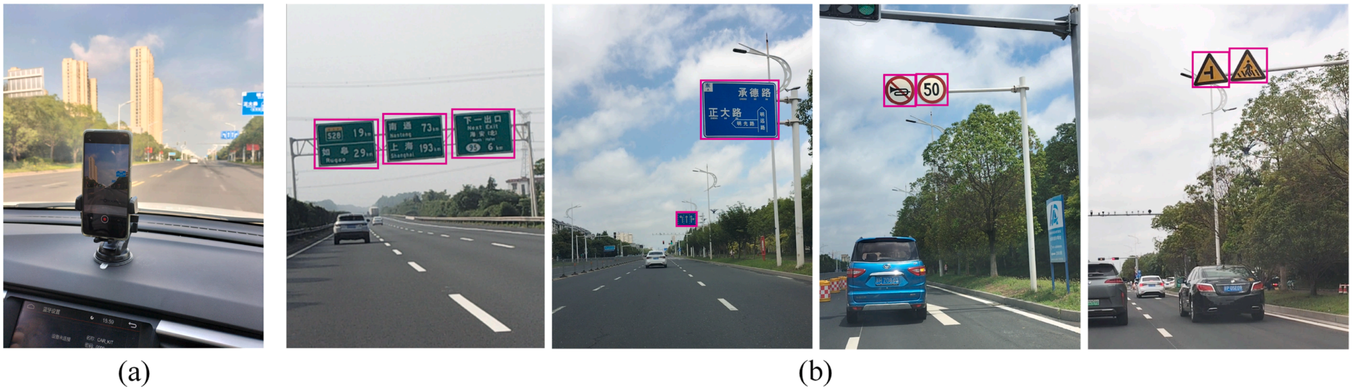


Fig. 6. Illustrations of (a) the device used for collecting the IVI-TSR dataset, and (b) the image examples and the annotations of the traffic sign instances in the IVI-TSR dataset.

generating the model parameters. For the detection head, since the bounding box regression branch requires the prediction results from the position classification branch to supervise the distributions of the anchors, we employed a "divide-and-combine" scheme to construct and optimize these two branches. Moreover, for the bounding box regression branch, the primarily regressed bounding boxes from the anchor regression terminal are used as the references to attain the instance-oriented aligned feature map for the determination of the final optimized bounding boxes. Hence, we also employed a "divide-and-combine" scheme to construct these two terminals. Concretely, first, we constructed the position classification branch alongside with the HRNet backbone based on the loss function in Eq. (1). During training, each training batch involved two images per GPU and optimized for 600 epochs. The learning rate was initially set as 0.001 and decayed to 0.0001 at the 401th epoch. When the position classification branch was constructed, we fixed the model parameters and further constructed the bounding box regression branch in an anchor-based manner. Here, only the anchor classification and regression terminals were constructed based on the loss functions in Eqs. (3) and (2), respectively. During training, these two terminals were trained for 400 epochs with a batch size of two images per GPU. The learning rate was configured as 0.001 initially and decreased to 0.0001 at the 201th epoch. Once these two terminals were constructed, we fixed all the model parameters and optimized the bounding box regression terminal for 200 epochs based on the loss function in Eq. (2). The learning rate was set invariably as 0.0001. Finally, we jointly optimized the entire SignHRNet for another 100 epochs with a constant learning rate of 0.0001.

4.3. Hyperparameter configuration

In the proposed SignHRNet, there is a hyperparameter K, which denotes the number of predesigned anchors deployed at each position of the foreground regions. The configuration of K might influence the traffic signs recognition accuracy and efficiency. Thus, to determine an optimal configuration for K, we conducted a set of experiments by considering different sizes and different aspect ratios of anchors. Specifically, we evaluated the cases with two sizes, three sizes, and four sizes of anchors and the cases with two aspect ratios, three aspect ratios, and four aspect ratios of anchors, thereby resulting in nine different combinations. The sizes of anchors were configured as 8, 16, 32, and 64, respectively, and the aspect ratios of anchors were configured as 0.4, 0.5, 1, and 2, respectively. The sizes and aspect ratios of anchors were manually determined based on the prior knowledge of the traffic signs in the images. For providing quantitative evaluations of these combinations on the traffic signs recognition performances, we employed the commonly used mean average precision (mAP) metric. The mAP evaluates the overall recognition performance by considering both the precision and recall values of all traffic sign categories. The quantitative evaluation results of these nine combinations are reported in Table 1. Specifically, mAP50 and mAP75, respectively, represent the mAP obtained at the IoU thresholds of 50 % and 75 % between the predicted bounding box and the ground-truth bounding box. Note that, all the experimental results in all the tables in this paper were reported with the average performance by repeating the same model for ten times with different randomly configured training and test sets. As reflected in Table 1, overall speaking, the mAP improved as either the number of sizes or the number of aspect ratios increased. It indicated that the increase of the number of anchors favored significantly to cover more shape and size variations of traffic signs, thereby effectively improving the recognition accuracy. Note that, when the number of sizes and the number of aspect ratios exceeded three, the recognition accuracy was almost stable. It meant that the combination of three sizes and three aspect ratios performed excellently to obtain a promising recognition accuracy. However, the increase of the number of anchors brought more computation burden due to the regressions and classifications of these anchors, thereby lowering the processing efficiency. Thus, by balancing the recognition accuracy and processing efficiency, we configured three sizes of [8, 16, 32] and three aspect ratios of [0.5, 1, 2], resulting in $K = 9$ anchors.

Next, we evaluated the impacts of the input image sizes on the traffic signs recognition accuracy and efficiency of the SignHRNet. To this end, we conducted a set of experiments by considering different configurations for the input image sizes. Specifically, for the TT-100K dataset, we set the input image sizes as 512 × 512 pixels, 1024 × 1024 pixels, and 2048 × 2048 pixels, respectively; for the CURE-TSD dataset, we set the input image sizes as 380 × 500 pixels, 760 × 1000 pixels, and 1236 × 1628 pixels, respectively; for the IVI-TSR dataset, we set the input image sizes as 500 × 375 pixels, 1000 × 750 pixels, and 2000 × 1500 pixels, respectively. The quantitative evaluation results of the SignHRNet with different input image sizes on the three test datasets are reported in Table 2. As reflected in Table 2, for each dataset, the traffic signs recognition accuracy improved as the input image size enlarged. It indicated that a larger input image exactly performed better than did a smaller input image. This is because the small-size traffic signs covered more image contents and exhibited more salient feature semantics in a larger input image, thereby improving the recognition accuracy of the small-size traffic signs. Note that, the SignHRNet with the largest input image sizes of 2048 × 2048 pixels, 1236 × 1628 pixels, and 2000 × 1500 pixels didn't achieve very significant recognition accuracy improvements compared with those of the SignHRNet with the medium input image sizes of 1024 × 1024 pixels, 760 × 1000 pixels, and 1000 × 750 pixels. However, the enlargement of the input image size would result in dramatic increase on the memory consumption and significantly lower the processing efficiency. Thus, by balancing the recognition accuracy and processing efficiency, we configured the input image sizes as 1024 × 1024 pixels, 760 × 1000 pixels, and 1000 × 750 pixels, respectively, for the TT-100K, CURE-TSD, and IVI-TSR datasets.

Finally, to examine the influences of the number of feature channels in each layer of the HRNet backbone on the traffic signs recognition performance, we also conducted a set of experiments by considering different configurations for the number of feature channels. Concretely, we tested the following six settings for the number of feature channels: 128, 256, 512, 768, 1024, and 1280. The quantitative evaluation results of the SignHRNet with different numbers of feature channels are reported in Table 3. Overall speaking, when the number of feature channels increased, the recognition accuracy improved accordingly. It demonstrated that more feature channels favored positively to extract stronger, more distinctive, and higher-quality feature semantics, which were beneficial to the recognition of the traffic signs under challenging scenarios, thereby promoting the overall recognition performance. Note that, when the number of feature channels exceeded 512, the recognition performance improvement was not very significant. Nevertheless, more feature channels would result in more network parameters and more memory consumptions, as well as significantly degrading the

Table 1
Quantitative evaluation results obtained by different combinations of sizes and aspect ratios of anchors.

Sizes Aspect ratios	[16,32] [0.5,1]	[16,32] [0.5,1,2]	[16,32] [0.4,0.5,1,2]	[8,16,32] [0.5,1]	[8,16,32] [0.5,1,2]	[8,16,32] [0.4,0.5,1,2]	[8,16,32,64] [0.5,1]	[8,16,32,64] [0.5,1,2]	[8,16,32,64] [0.4,0.5,1,2]
mAP (%)	55.93	65.87	66.77	69.94	72.85	72.86	70.75	72.87	72.87
mAP50 (%)	80.47	90.36	91.32	94.27	96.48	96.48	95.01	96.49	96.49
mAP75 (%)	67.76	78.61	79.59	83.09	85.31	85.32	83.85	85.33	85.33

Table 2
Quantitative evaluation results obtained by different input image sizes.

Dataset	TT-100K			CURE-TSD			IVI-TSR		
Image size (pixels)	512 × 512	1024 × 1024	2048 × 2048	380 × 500	760 × 1000	1236 × 1628	500 × 375	1000 × 750	2000 × 1500
mAP (%)	69.54	73.72	73.95	69.73	74.05	74.21	66.53	70.78	70.96
mAP50 (%)	93.77	97.08	97.31	93.92	97.24	97.44	91.16	95.13	95.45
mAP75 (%)	82.68	85.97	86.22	82.85	86.23	86.41	79.97	83.73	83.97

Table 3
Quantitative evaluation results obtained by different numbers of feature channels.

Number of feature channels	128	256	512	768	1024	1280
mAP (%)	65.56	70.33	72.85	73.09	73.11	73.12
mAP50 (%)	90.12	94.71	96.48	96.74	96.77	96.79
mAP75 (%)	78.43	83.46	85.31	85.53	85.55	85.56

processing efficiency. Thus, by balancing the recognition accuracy and processing efficiency, we configured the number of feature channels as 512.

4.4. Traffic signs recognition

At the test stage, the aforementioned three datasets were used to examine the traffic signs recognition performance of the proposed SignHRNet. For a test image fed into the SignHRNet, the outputs predicted by the bounding box regression branch constituted the detected bounding boxes of the traffic sign instances in this image, and the category information corresponding to each traffic sign instance can be attained from the outputs predicted by the position classification branch. To provide quantitative evaluations on the traffic signs recognition results, we also employed the mAP metric.

Table 4
Quantitative evaluation results obtained by different models on the three test datasets.

Model	Dataset	mAP (%)	mAP50 (%)	mAP75 (%)	FPS
SignHRNet	TT-100K	73.72	97.08	85.97	28
	CURE-TSD	74.05	97.24	86.23	
	IVI-TSR	70.78	95.13	83.73	
	Overall	72.85	96.48	85.31	
Dense-RefineDet	TT-100K	62.25	91.17	78.64	17
	CURE-TSD	53.84	86.89	73.51	
	IVI-TSR	56.27	88.76	75.47	
	Overall	57.45	88.94	75.87	
GMAPNet	TT-100K	67.23	93.40	81.54	34
	CURE-TSD	70.84	95.15	83.77	
	IVI-TSR	64.37	92.33	79.96	
	Overall	67.48	93.63	81.76	
TSingNet	TT-100K	67.54	93.51	81.72	23
	CURE-TSD	65.11	92.51	80.39	
	IVI-TSR	65.52	92.63	80.62	
	Overall	66.06	92.88	80.91	
MF-SSD	TT-100K	62.44	91.22	78.75	36
	CURE-TSD	55.06	87.28	74.13	
	IVI-TSR	59.21	89.57	76.69	
	Overall	58.90	89.36	76.52	
SegU-Net	TT-100K	65.06	92.50	80.37	19
	CURE-TSD	69.57	94.60	83.01	
	IVI-TSR	63.87	91.72	79.61	
	Overall	66.17	92.94	81.00	
MFPSANet	TT-100K	69.91	94.73	83.25	25
	CURE-TSD	70.76	95.13	83.71	
	IVI-TSR	66.33	92.83	80.92	
	Overall	69.00	94.23	82.63	
Mask R-CNN	TT-100K	64.66	92.37	80.11	10
	CURE-TSD	58.09	89.16	76.05	
	IVI-TSR	62.24	91.15	78.57	
	Overall	61.66	90.89	78.24	
Enhance-Net	TT-100K	64.98	92.48	80.32	16
	CURE-TSD	62.13	91.13	78.53	
	IVI-TSR	62.86	91.47	79.08	
	Overall	63.32	91.69	79.31	
Faster R-CNN	TT-100K	73.68	97.11	85.93	10
	CURE-TSD	74.08	97.29	86.24	
	IVI-TSR	70.77	95.17	83.71	
	Overall	72.84	96.52	85.29	

Table 4 reports the quantitative assessment results based on the mAP metric on the three test datasets by using the proposed SignHRNet. Specifically, the results of mAP50 and mAP75 were also listed. The precision-recall (PR) curves on the three datasets are shown in Fig. 7. As reflected in Table 4 and Fig. 7, the traffic sign instances in the test images were excellently detected and recognized with correct categories in spite of the considerable variations of the scenes and the great diversities of the traffic sign instances. The background objects were reasonably distinguished from the traffic sign instances with a quite low false recognition rate with regard to the precision indicator. Moreover, the traffic sign instances in the test images were promisingly located and identified with a quite low missing detection rate with respect to the recall indicator. Thus, the proposed SignHRNet performed effectively on the three test datasets towards traffic signs recognition. To be specific, on the TT-100K dataset, the quantitative assessment results are about 73.72 %, 97.08 %, and 85.97 % for the mAP, mAP50, and mAP75 metrics, respectively. The recognition performance with the mAP, mAP50, and mAP75 metrics is about 74.05 %, 97.24 %, and 86.23 %, respectively, on the CURE-TSD dataset. For the IVI-TSR dataset, the SignHRNet obtained a recognition accuracy with an mAP of 70.78 %, an mAP50 of 95.13 %, and an mAP75 of 83.73 %, respectively. Noteworthily, the precision indicator has a higher value than that of the recall indicator on the TT-100K and CURE-TSD datasets, which means that the SignHRNet generated a small portion of false alarms, while a non-negligible portion of traffic sign instances were failed to be completely identified. The missing recognitions were primarily caused by the traffic sign instances with remarkably challenging conditions, such as small sizes due to long vision distances, blurs caused by sensor shakes or weather conditions, low contrasts due to illumination conditions, deformations due to large perspectives, and occlusions caused by neighboring targets. These issues degraded the semantic saliencies of the

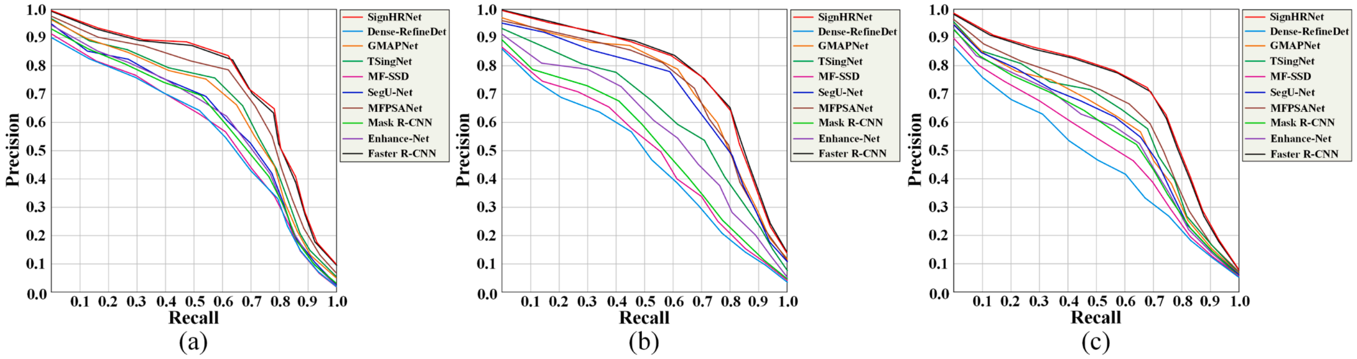


Fig. 7. Precision-recall curves of different models on (a) the TT-100K dataset, (b) the CURE-TSD dataset, and (c) the IVI-TSR dataset.

traffic sign instances in the resultant feature maps, thereby leading to the omissions and the degradation of the recall indicator. In contrast, on the IVI-TSR dataset, the value of the recall indicator is higher than that of the precision indicator, which means that more background objects were incorrectly characterized and categorized as the traffic signs. The false recognitions were mainly caused by the roadside advertising boards that exhibited extremely similar geometric and semantic properties to the traffic signs. As a result, these background objects exhibited strong semantic presences in the resultant feature maps, thereby leading to the misidentifications and the lowering of the precision indicator. Comparatively, the best performance fell on the CURE-TSD dataset and a relatively worse accuracy appeared on the IVI-TSR dataset due to its more challenging background scenarios and more complicated self-conditions of the traffic sign instances. Overall speaking, the proposed SignHRNet demonstrated competitive accuracies and feasible solutions on the three test datasets for the task of traffic signs localization and recognition. As statistics, the average mAP, average mAP50, and average mAP75 are about 72.85 %, 96.48 %, and 85.31 %, respectively, by treating the three test datasets as a whole. Although the recognition accuracy is quite promising and competitive, it still remains a certain gap in comparison with the ideal human-level qualities and accuracies that allow no false recognitions and missing detections.

The complexities and challenges that impacted the recognition performance in the three test datasets resided in the following cases. (1) Some traffic sign instances had extremely small sizes covering only dozens of or hundreds of pixels in the test images owing to the long image capturing distances. (2) Some traffic sign instances were severely deformed and squished due to the large shooting perspectives, especially the unidentifiable signals on the sign boards. (3) Some traffic sign instances were back on to the sensors while conducting image collection, thereby resulting in the signal unavailability and pattern discrepancy. (4) Some traffic sign instances exhibited with different-level rotations in the test images caused by the sensor rotations, particularly the rectangular traffic signs. (5) The traffic sign instances showed different shapes for their special uses and meanings, such as round shapes, rectangular shapes, triangular shapes, octagonal shapes, etc. (6) Different categories of traffic signs were painted with different colors (e.g., blue, red, green, brown, yellow, etc.) and contained different-form contents (e.g., figures and characters). (7) Some traffic sign instances suffered from different-level occlusions caused by the neighboring background objects, thereby leading to the geometric structure and signal content incompleteness. (8) Some traffic sign instances exhibited with low contrasts with their surrounding environments due to the illumination conditions, such as dim lights and bright lights. (9) Some traffic sign instances underwent different-level blurs, which were caused by the sensor shakes or bad weather conditions, such as rain and fog. (10) The dense distributions and close connections of some traffic sign instances brought difficulties in accomplishing clear separations, especially the rectangular traffic signs. (11) Some background objects, such as advertising boards, showed quite similar pattern and texture attributes to the traffic signs. (12) The traffic sign instances existed in complicated scenes, such as road scenarios and highway scenarios. (13) The CURE-TSD dataset suffered from the manually added different types of effects and different levels of challenges. The aforementioned cases in the test images exactly resulted in the feature quality and distinguishability degradations of the traffic sign instances in the resultant feature maps, thereby unquestionably impeding the precise localization and correct identification of the traffic signs, probably leading to the failures in guaranteeing the detection integrity to achieve a high recall rate or the faults in introducing the background targets or misclassifying a traffic sign into an incorrect category to lower the recognition rate. Nevertheless, as reflected by the traffic signs recognition results in Table 4 and Fig. 7, the proposed SignHRNet still behaved excellently with quite high precision and recall evaluations on the handling of the traffic sign instances of diverse self-conditions under varying environments. The meritorious performance of the SignHRNet was embodied in the following aspects. Firstly, stacked by an HRNet structure as the feature extraction backbone to extract feature representations in different-size subspaces, the SignHRNet performed superiorly in providing high-quality, strong-semantic feature encodings at each branch, which was beneficial to the localization and recognition of traffic signs with different sizes, varying conditions, and diverse scenarios. Secondly, embedded with the dual-attention module for feature semantic recalibration by comprehensively considering the channel and spatial feature significances, the SignHRNet was further boosted to attend to semantic-important and task-aware features, thereby effectively promoting the feature representation quality and robustness. Last but not least, designed with the semi-anchoring guided strategy and the feature alignment module, the SignHRNet can efficiently locate the traffic sign regions and accurately determine the tight bounding boxes of the traffic signs, thereby promoting the detection and recognition efficiencies and accuracies.

To visually check the performance of the proposed SignHRNet in handling the small-size traffic signs and the traffic signs under complicated scenarios and weather conditions, Figs. 8, 9, and 10 also present some sample test images, alongside with the traffic signs recognition results, from the three test datasets. Apparently, the traffic sign instances showing varying sizes, shapes, colors, and patterns, exhibiting different contrasts, rotations, and deformations, contaminated by different types of effects and weather conditions, existing in different road scenarios, suffering from occlusions, and distributing densely were promisingly identified and located with tight bounding boxes. Specifically, as shown in these figures, some traffic sign instances occupy quite small image contents and exhibit very small sizes in the images due to the long shooting distances and some traffic sign instances undergo severe geometric deformations due to the large shooting perspectives. As a matter of fact, accurately identifying these traffic sign instances is tough because of the lack of sufficient feature presences or the lack of representative feature semantics. Furthermore, in some images, the traffic sign instances demonstrate quite low or high brightness due to the variations of illumination conditions. Some images in the IVI-TSR

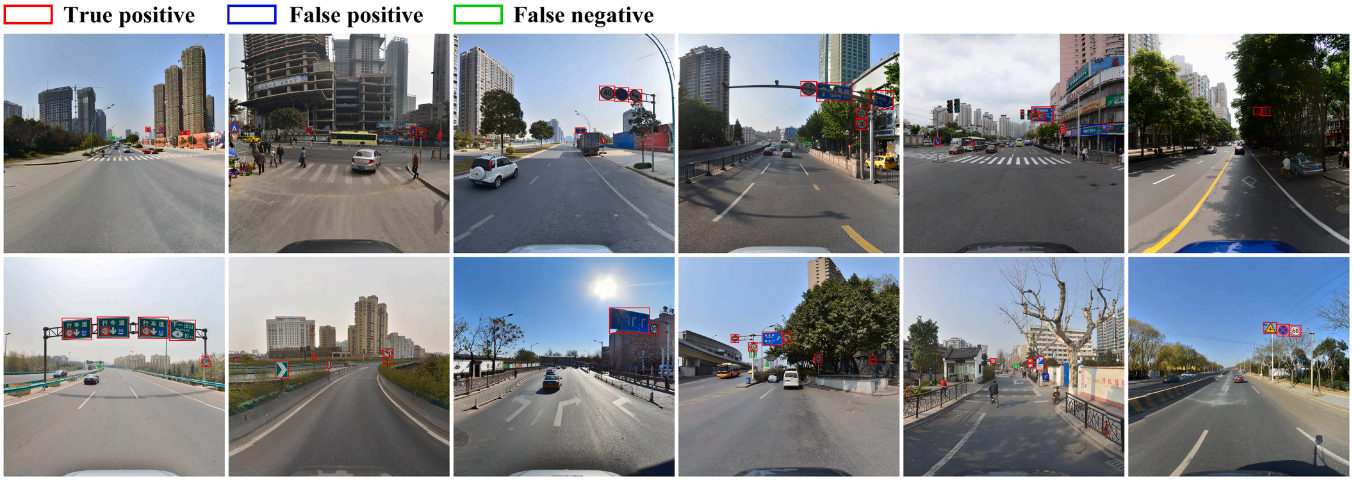


Fig. 8. Traffic signs recognition results on the TT-100K dataset.

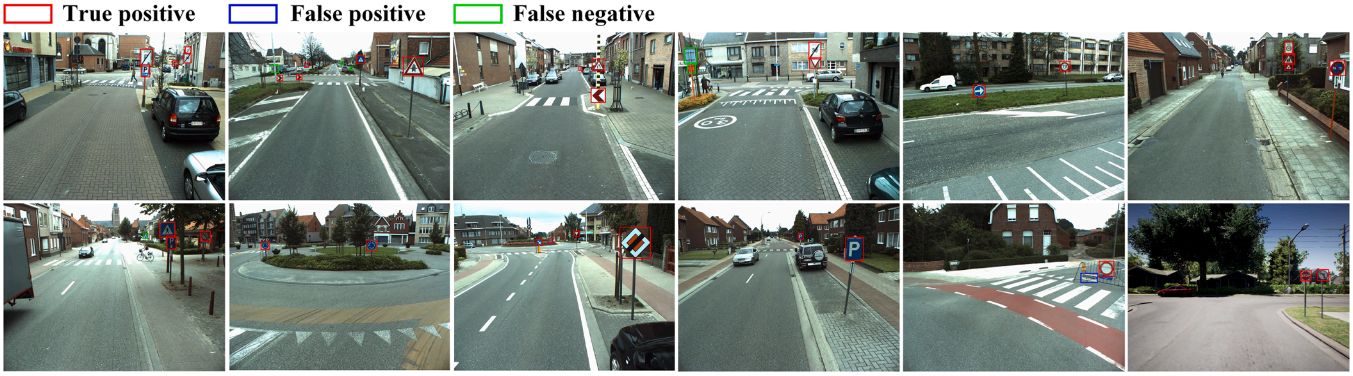


Fig. 9. Traffic signs recognition results on the CURE-TSD dataset.

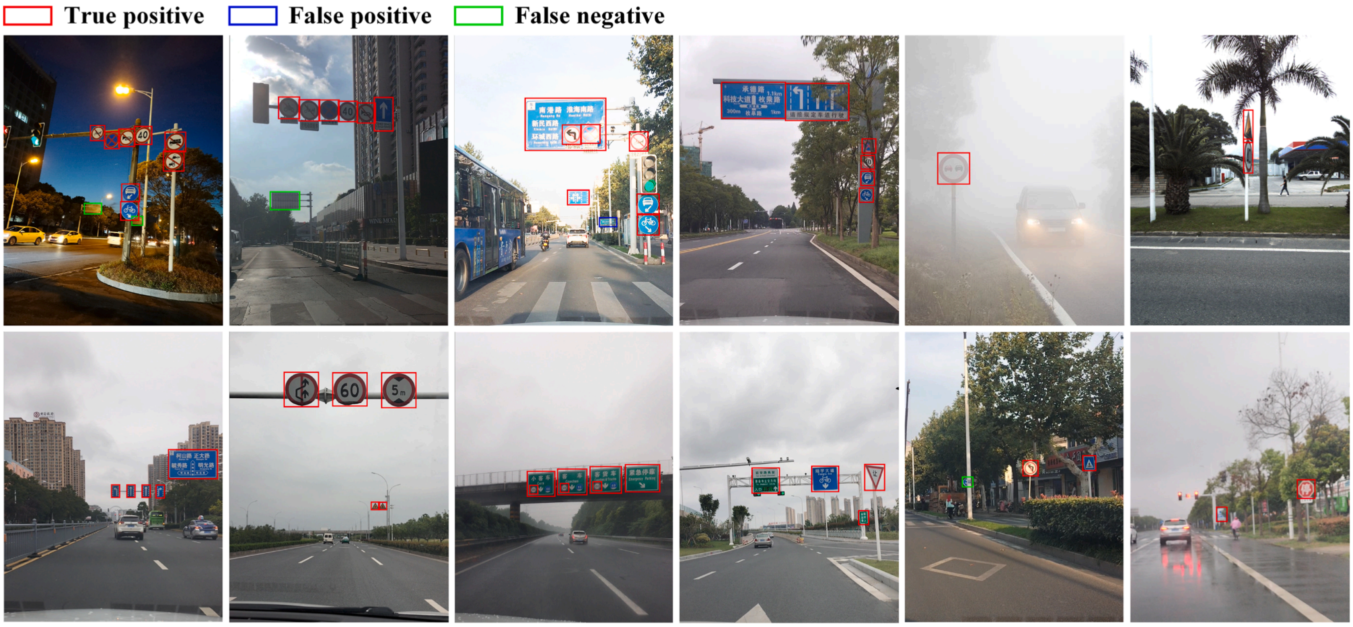


Fig. 10. Traffic signs recognition results on the IVI-TSR dataset.

dataset were even collected at dusk or early night. These traffic sign instances are easily to be regarded as the background components due to their non-salient features. Moreover, some traffic sign instances in the test images are contaminated by bad weather conditions (e.g., rain and fog), manually added effects (e.g., the images in the CURE-TSD dataset), and image blurs, which lead to detail unsharpnesses and low visibilities.

As a result, their feature saliency and distinguishability are also impacted to some extent. In addition, the nearby vegetation and the other in-sight objects shield parts of the traffic sign boards, thereby resulting in the structure incompleteness and content loss. This is also a challengeable issue to precisely and completely recognize these traffic sign instances as they might be less concentrated on and less intensively characterized. Last but not least, the traffic sign instances exhibited different levels of rotations, different patterns of spatial distributions, different shapes, and different color appearances under different complicated road scenarios (e.g., urban roads, suburban roads, and highways) in the images. Some images even contained very densely distributed or overlapped traffic sign instances. It is not easy to accurately and tightly determine the bounding boxes of these traffic sign instances due to the interference of the background. Fortunately, benefitting from the employment of the HRNet architecture as the feature extraction backbone to maintain a high-resolution branch through the entire network and repeatedly carry out cross-branch feature exchange, the boosting of the dual-attention module for channel and spatial feature semantic promotions, and the assistance of the feature alignment module for instance-oriented feature generation, the traffic sign instances under the aforementioned challenging conditions were nicely located and recognized.

However, as shown in Fig. 11, some traffic sign instances have extremely small sizes occupying only dozens of pixels and containing indiscernible signal details (Fig. 11(a) and (d)). Some traffic sign instances are severely occluded with only a very small visible part (Fig. 11 (b)). Some traffic sign instances are back on to the sensors with no recognizable attributes (Fig. 11(c)). As a result, owing to the insufficient, non-salient, and indistinguishable feature semantics, these traffic sign instances were not successfully identified. In addition, some roadside advertising boards have extremely similar structure and content properties to the traffic signs, especially the rectangular advertising boards, and exhibited very strong feature presences in the resultant feature maps (Fig. 11(d)). Hence, they were incorrectly recognized as the traffic signs.

At the test stage, the processing time was also recorded to evaluate the efficiency of the proposed SignHRNet. Specifically, the processing efficiency was measured by the frames per second (FPS) indicator, which denoted the number of image frames being processed each second. On average, the proposed SignHRNet achieved a processing efficiency of about 28 FPS.

4.5. Comparative analyses

On the purpose of further analyzing the feasibility and superiority of the designed SignHRNet, a group of intensive comparative experiments were also conducted with some recently developed state-of-the-art deep learning models that served for traffic signs recognition tasks. The selected models include: Dense-RefineDet (Sun et al., 2020), group multiscale attention pyramid network (GMAPNet) (Shen et al., 2021), scale-aware and context-rich feature network (TSingNet) (Liu et al., 2021), multi-feature SSD (MF-SSD) (Jin et al., 2020), SegU-Net (Kamal et al., 2020), multiscale fusion and prime sample attention network (MFPSANet) (Cao et al., 2021), mask R-CNN (Serna and Ruichek, 2020), and prior enhancement network (Enhance-Net) (Ahmed et al., 2022). All the models kept the same network architectures as those proposed in the corresponding papers without any modifications on their network architectures. Concretely, the Dense-RefineDet and the MF-SSD adopted the VGG-16 as the backbone, the GMAPNet adopted the ResNet-50 as the backbone, the TSingNet adopted the ResNet-50 with a bilateral feature pyramid network (FPN) architecture as the backbone, the SegU-Net and the Enhance-Net adopted the SegU-Net architecture as the backbone, the MFPSANet adopted the HRNet-W18 as the backbone, and the mask R-CNN adopted the ResNet-101 with an FPN architecture as the backbone. Among the eight models, the first four models are one-stage object detection architectures and the other four models follow the two-stage object detection pipelines. Specifically, all the models investigated the multiscale or multilevel feature semantics and reasonably combined them to provide high-quality and robust instance encodings, such as skip connections and feature pyramid structures. Worth mentioning, similar to our proposed SignHRNet, the MFPSANet also employed the HRNet formulation as the feature extraction backbone. In addition, feature attention mechanisms and context augmentation schemes were also considered in some models for boosting the feature semantic quality. Aiming at providing comparative analyses on the same baseline, all the eight models were trained and tested on the three datasets used in this paper with the same training and test data partition scheme and the same input sizes of the images as those of the SignHRNet. The performances of these models quantitatively evaluated by the means of mAP, mAP50, and mAP75, as well as their processing efficiencies measured by the FPS indicator, are reported in detail in Table 4. The PR curves of these models are shown in Fig. 7.

As shown by the quantitative statistics in Table 4, the MFPSANet, GMAPNet, SegU-Net, and TsingNet behaved similarly, but showed obviously more advantageous recognition accuracies than the other models, while the recognition performances of the MF-SSD and Dense-RefineDet were relatively lower. Interestingly, the one-stage model GMAPNet even performed better than the three two-stage models SegU-Net, Enhance-Net, and mask R-CNN. Specifically, the accuracy difference between the MFPSANet and the Dense-RefineDet was about 11.55 % with respect to the average mAP on the three test datasets. Such a performance superiority of the MFPSANet benefitted from the HRNet architecture with an attention sample selection strategy, which comprehensively fused multilevel and multiscale feature semantics, and attended to significant feature encodings. As a result, the strong feature semantics well supported the detection of the small-size traffic signs. In contrast, the GMAPNet took advantage of the pyramidal feature representations and leveraged a group-wise multiscale attention scheme, whereas the TSingNet also employed a pyramidal feature extractor and

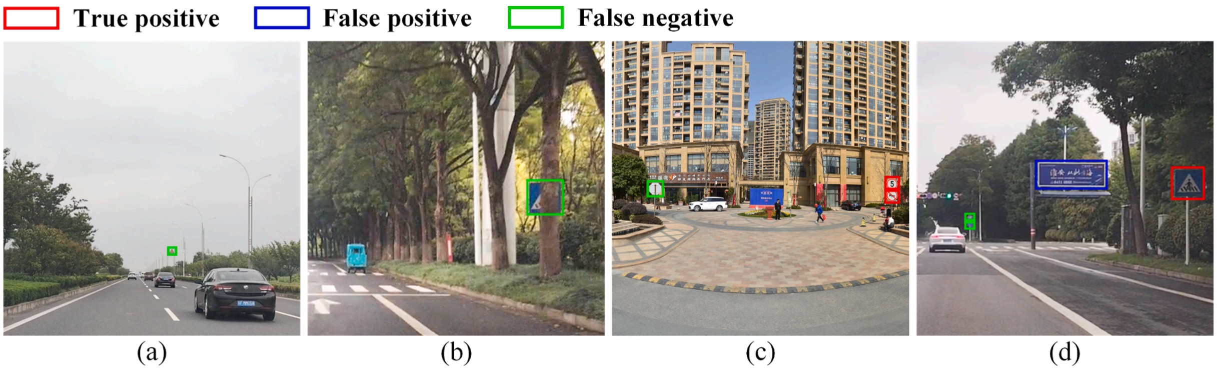


Fig. 11. Illustrations of (a) an extremely small-size traffic sign, (b) a severely occluded traffic sign, (c) a traffic sign back on to the sensor, and (d) an advertising board.

exploited scale-aware feature representations and multiscale contextual feature contents. Therefore, the extracted features of these two models were also semantically strong and instance-sensitive, which served positively to locate small-size instances and the instances having complex surroundings and varying self-conditions. The performance declines of the MF-SSD and the Dense-RefineDet were primarily caused by the output lower-resolution feature maps used for traffic signs classification and bounding boxes regression. Thus, they showed weak capabilities in handling the small-size traffic signs.

Comparatively, the proposed SignHRNet demonstrated significant improvement over all the compared models. For example, the accuracy difference between the SignHRNet and the best model MFPSANet was about 3.85 % with regard to the average mAP metric. Note that, the accuracy difference between the SignHRNet and the Dense-RefineDet was even about 15.40 % with regard to the average mAP metric. Specifically, the performance gains of the SignHRNet over the mask R-CNN benefitted significantly from the HRNet backbone with the dual-attention module for higher-quality feature semantics extraction. Through comparative analyses, we concluded that the proposed SignHRNet supplied a competitive, reliable, and effective solution to highly accurate traffic signs recognition applications.

As another comparative experiment, we further analyzed the efficiency and effectiveness between the proposed semi-anchoring guided architecture and the RPN-based faster R-CNN architecture (Ren et al., 2017). To this end, we replaced the feature extraction backbone in the faster R-CNN by the HRNet feature extraction backbone with the dual-attention module proposed in this paper. That is, the modified faster R-CNN had the same feature extraction backbone as the SignHRNet, but having different detection heads. Specifically, the modified faster R-CNN employed the anchor-based architecture, while the SignHRNet employed the semi-anchoring guided architecture. Likewise, the modified faster R-CNN was trained and tested on the three datasets with the same training and test data partition scheme and the same input sizes of the images as those of the SignHRNet. The performance of the modified faster R-CNN model quantitatively evaluated by the means of mAP, mAP50, and mAP75, as well as its processing efficiency measured by the FPS indicator, are reported in detail in Table 4. The PR curves are shown in Fig. 7. As reflected in Table 4, the modified faster R-CNN achieved equal-matched traffic signs recognition performance compared with the proposed SignHRNet. However, it showed significantly lower processing efficiency compared with the SignHRNet. Thus, it demonstrated that the proposed semi-anchoring guided architecture can achieve compatible accuracy with the anchor-based architectures, but achieving significantly higher processing efficiency than the anchor-based architectures.

4.6. Ablation studies

In the proposed SignHRNet, the high-resolution branch of the HRNet backbone, the dual-attention module, and the feature alignment module contributed positively and significantly to the promotions of the feature representation quality and the quality of the regressed bounding boxes. As ablation studies, we intently examined the advanced superiorities of these three modules to the enhancement of the traffic signs recognition accuracies. To realize this objective, first, we conducted a group of ablation experiments to evaluate the dual-attention module. To be specific, first, we removed all the dual-attention modules from the HRNet backbone to abolish the feature semantic attention mechanisms. The modified architecture was named as the SignHRNet-NULL. Second, we removed the spatial-specific attention unit from the dual-attention module, leaving only the channel-specific attention unit for recalibrating channel features. The modified architecture was named as the SignHRNet-CSA. Third, we removed the channel-specific attention unit from the dual attention module, leaving only the spatial-specific attention unit for recalibrating spatial features. The modified architecture was named as the SignHRNet-SSA. Fourth, we integrated the dual-attention module at the beginning of each stage to perform feature attention, i.e. at the point after cross-branch feature exchange in each stage. The modified architecture was named as the SignHRNet-BA. Finally, we integrated the dual-attention module in the middle of each stage to perform feature attention. The modified architecture was named as the SignHRNet-MA.

Table 5 details the traffic signs recognition results obtained by these modified models on the three test datasets. Likewise, the mAP metric, as well as the mAP50 and mAP75, were also leveraged for quantitative analyses and comparisons. Note that, by abandoning the dual-attention module for feature semantic promotion, the SignHRNet-NULL behaved less promisingly with considerable accuracy degradations on all the three test datasets. The reason is that, without the dual-attention module for attending to the important, informative channel features and the

Table 5
Quantitative evaluation results of different modified models on the three test datasets.

Model	Dataset	mAP (%)	mAP50 (%)	mAP75 (%)	FPS
SignHRNet-NULL	TT-100K	68.66	94.45	82.61	31
	CURE-TSD	70.32	94.87	83.48	
	IVI-TSR	65.14	92.51	80.29	
	Overall	68.04	93.94	82.13	
SignHRNet-CSA	TT-100K	70.54	95.83	83.87	29
	CURE-TSD	72.18	95.97	84.18	
	IVI-TSR	66.88	93.66	81.34	
	Overall	69.86	95.15	83.13	
SignHRNet-SSA	TT-100K	70.22	95.51	83.49	29
	CURE-TSD	71.94	95.63	83.82	
	IVI-TSR	66.61	93.40	81.07	
	Overall	69.59	94.85	82.79	
SignHRNet-BA	TT-100K	73.44	96.83	85.73	28
	CURE-TSD	73.71	97.02	85.98	
	IVI-TSR	70.45	94.89	83.47	
	Overall	72.53	96.25	85.06	
SignHRNet-MA	TT-100K	73.23	96.67	85.51	28
	CURE-TSD	73.49	96.84	85.77	
	IVI-TSR	70.28	94.66	83.25	
	Overall	72.33	96.06	84.84	
SignHRNet-lower	TT-100K	68.74	94.55	82.68	33
	CURE-TSD	70.37	94.92	83.54	
	IVI-TSR	65.28	92.63	80.36	
	Overall	68.13	94.03	82.19	
SignHRNet-wo-FAM	TT-100K	72.94	96.53	85.27	28
	CURE-TSD	73.17	96.71	85.56	
	IVI-TSR	69.03	94.07	82.46	
	Overall	71.71	95.77	84.43	
SignHRNet-FGS	TT-100K	73.73	97.09	85.97	22
	CURE-TSD	74.05	97.23	86.24	
	IVI-TSR	70.79	95.14	83.74	
	Overall	72.86	96.49	85.32	

salient, task-oriented spatial features, the semantic and quality of the extracted feature maps were weakened and lowered, thereby leading to the performance decline on the cases of the traffic sign instances with challenging conditions, such as the small-size traffic signs. Overall, the accuracy decline of the SignHRNet-NULL with regard to the mAP was about 4.81 % on the three test datasets compared with that of the SignHRNet. Furthermore, by integrating the channel-specific attention unit or the spatial-specific attention unit, the performances of the SignHRNet-CSA and SignHRNet-SSA improved significantly compared with that of the SignHRNet-NULL. This well demonstrated that each of the two attention units behaved positively to the promotion of the feature semantic quality, thereby leading to the improvement of the recognition accuracy. Comparatively, the SignHRNet-CSA achieved slightly higher mAP than the SignHRNet-SSA, which means that the channel-specific attention unit is relatively powerful than the spatial-specific attention unit in feature semantic recalibration. For visual comparisons, Fig. 12 show some examples of the feature saliency maps of these models generated in Branch 1 with and without feature attention mechanisms. These feature saliency maps were generated and visualized using the CNN-based Grad-CAM tool (Selvaraju et al., 2017), which used the gradients of the traffic sign classes to produce a localization map highlighting the important regions in the image for predicting the traffic signs. That is, the brighter the regions, the more important and more salient the feature semantics in the regions. It can be observed that, with feature attention mechanisms, the feature saliencies focused more and better on the traffic sign regions, especially for the small-size traffic signs, which were well located and whose feature semantics were well highlighted, thereby improving the recognition accuracy of the small-size traffic signs. In addition, by integrating the dual-attention module at different positions (i.e. at the beginning and in the middle) in each stage, the recognition performances of the SignHRNet-BA and SignHRNet-MA were slightly degraded. However, the degradations were not significant compared with that of the SignHRNet. This was mainly caused by the quality of the feature semantics used in the cross-branch feature exchange procedure. By mounting the dual-attention module at the end of each stage, the feature semantics from different branches were first promoted and then integrated, which was beneficial to generate higher-quality feature representations in each branch, thereby improving the recognition accuracy. In conclusion, the design pattern of the dual-attention module meant significantly to the feature semantic enhancement.

Except for the dual-attention module contributing to the recognition accuracy improvement of the small-size traffic signs, we further evaluated the high-resolution branch (i.e. Branch 1) of the HRNet backbone to the recognition accuracy improvement of the small-size traffic signs. To this end, we discarded the feature map generated in Branch 1 and only applied the feature maps generated in the last three lower-resolution branches to conduct traffic signs recognition. The modified architecture was named as the SignHRNet-lower. As reported in Table 5, the recognition accuracies of the SignHRNet-lower declined significantly on the test datasets compared with those of the SignHRNet. The performance degradation was mainly caused by the missing recognition of the small-size traffic signs, which showed quite low feature saliencies in the lower-resolution feature maps. To provide visual comparisons, Fig. 13 also shows some sample results obtained by the SignHRNet-lower and the SignHRNet. As reflected in Fig. 13, some small-size traffic signs failed to be recognized by the SignHRNet-lower, whereas they were successfully recognized by the SignHRNet. In conclusion, it demonstrated that by maintaining the high-resolution branch across the entire network, it is significantly beneficial to the recognition of the small-size traffic signs.

Next, we removed the feature alignment module and its associated layers from the detection head, and directly used the outputs exported by the anchor regression terminal as the predicted bounding boxes. The modified architecture was named as the SignHRNet-wo-FAM. As reported in Table 5, the recognition accuracy decline also appeared on the SignHRNet-wo-FAM without the connection of the feature alignment module. However, the decline was not as dramatic as that of the SignHRNet-NULL. To be specific, the accuracy decline was only about 1.14 % with respect to the mAP. The main adverse impact was the quality degradation of the predicted bounding boxes of the traffic signs, especially those traffic sign instances with small sizes, vague boundaries, or partial occlusions. As shown in Fig. 14, the SignHRNet (Fig. 14(b)) generated more accurate and tighter bounding boxes than the SignHRNet-wo-FAM (Fig. 14(a)). For clear visual comparisons, Fig. 14 (d) and (c) also show the feature saliency maps generated with and without the feature alignment module. Obviously, the feature saliency map generated with the feature alignment module focuses more tightly on the traffic sign regions. In conclusion, we confirmed that the feature alignment module behaved meaningfully and usefully to the enhancement of the recognition performance.

As another ablation experiment to analyze the effectiveness and efficiency of the semi-anchoring guided strategy, we modified the detection head of the SignHRNet into the standard full-anchoring guided strategy. Concretely, we cut the connection between the position classification branch and the bounding box regression branch, and deployed densely a set of K predefined anchors at each position of the multi-task feature map. In this way, each position of the feature map had to carry out traffic sign bounding boxes regression without any supervisions about the traffic sign regions. The modified architecture was named as the SignHRNet-FGS. As reported in Table 5, the SignHRNet-FGS achieved equal-matched performance to that of the SignHRNet. It meant that the semi-anchoring guided strategy with less number of anchors can still act compatibly with the full-anchoring guided strategy, thereby proving the effectiveness of the proposed semi-anchoring guided strategy. However, as reflected by the processing speed, the SignHRNet-FGS attained an FPS of 22, which was slower than that of the SignHRNet. It indicated that the SignHRNet operated more efficiently than the SignHRNet-FGS. The efficiency decline of the SignHRNet-FGS was caused by the regressions of the large-volume dense anchors at all the positions of the feature map, whereas the SignHRNet only required to regress a small number of anchors located in the traffic sign regions,

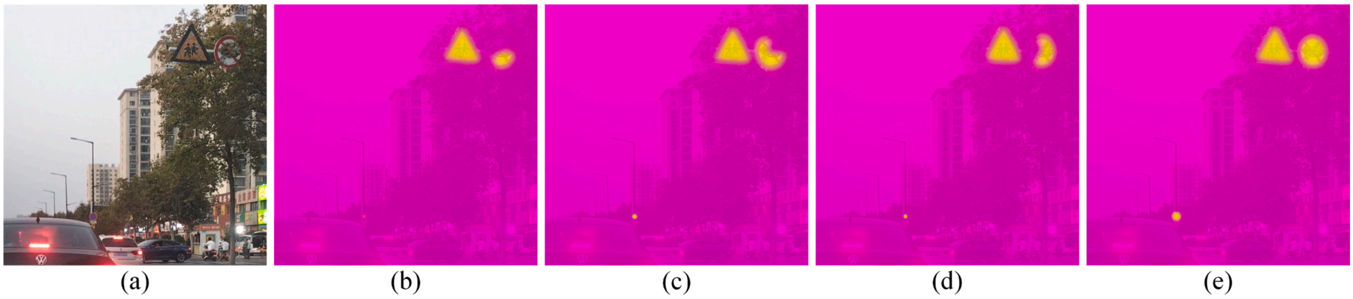


Fig. 12. (a) Test image and feature saliency maps generated (b) without the dual-attention module, (c) with only the channel-specific attention unit, (d) with only the spatial-specific attention unit, and (e) with the dual-attention module.

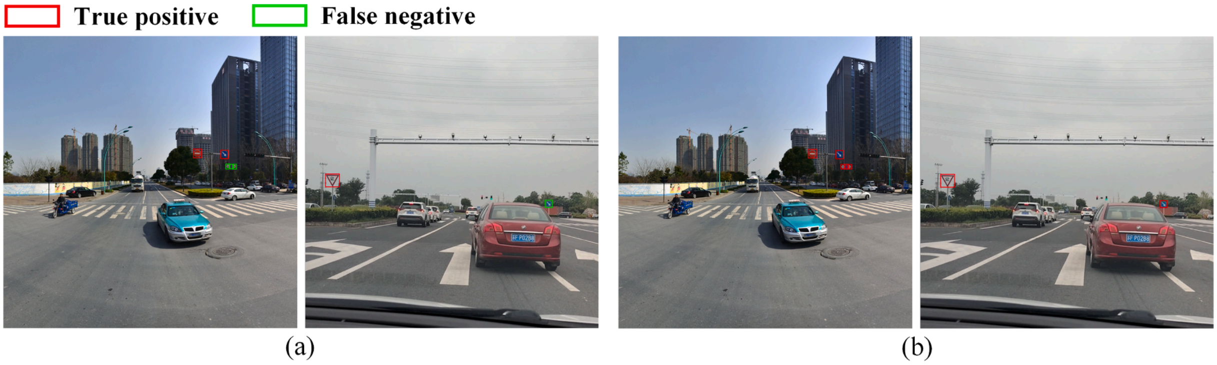


Fig. 13. Traffic signs recognition results obtained by (a) the SignHRNet-lower and (b) the SignHRNet.

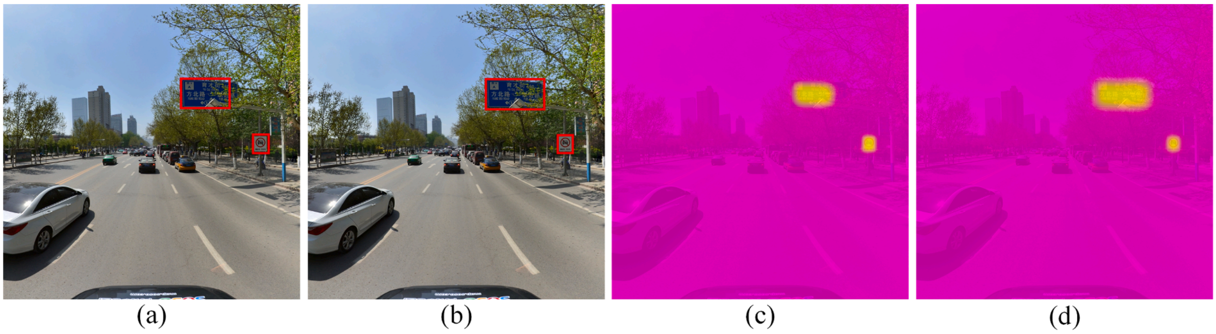


Fig. 14. Predicted bounding boxes by (a) the SignHRNet-wo-FAM and (b) the SignHRNet. Feature saliency maps generated (c) without and (d) with the feature alignment module.

thereby significantly improving the processing efficiency. In conclusion, the semi-anchoring guided strategy performed both effectively and efficiently.

Aiming at analyzing the superiorities of the proposed feature attention module in promoting the feature semantics, we also conducted a set of ablation experiments to make a comparison with the existing feature attention mechanisms. The following four attention mechanisms were considered: SE block (Hu et al., 2020), coordinate attention (Hou et al., 2021), CBAM (Woo et al., 2018), and DA module (Fu et al., 2019). To be specific, we substituted the proposed feature attention module in the SignHRNet with the SE block, coordinate attention, CBAM, and DA module, respectively, to construct four modified architectures. These modified architectures were named as the SignHRNet-SE, SignHRNet-CA, SignHRNet-CBAM, and SignHRNet-DA, respectively. As reported in Table 6, the SignHRNet-DA achieved the best recognition accuracy among the four modified models, whereas the SignHRNet-CBAM behaved less promisingly. Besides, the SignHRNet-CA performed slightly better than the SignHRNet-SE due to the embedding of the position information when exploiting the channel feature saliencies. The performance gains of the SignHRNet-DA benefitted from the concurrent consideration and integration of both the channel and spatial feature significances, thereby effectively promoting the feature representation quality. Nevertheless, the SignHRNet with the proposed feature attention module demonstrated significant accuracy improvement compared with these modified models, which convinced the effectiveness of the designed feature attention strategy. Note that, as reported by the processing speeds, the SignHRNet-DA showed lower efficiency than the other modified models due to the complex matrix operations in the DA module, whereas the SignHRNet-SE behaved quite efficiently. However, the proposed SignHRNet showed obvious efficiency improvement compared with the SignHRNet-DA and compatible efficiency with the SignHRNet-CA due to the lightweight architecture of the proposed feature attention module.

Table 6
Quantitative evaluation results of the modified models with different feature attention mechanisms on the three test datasets.

Model	Dataset	mAP (%)	mAP50 (%)	mAP75 (%)	FPS
SignHRNet-SE	TT-100K	69.44	94.95	83.36	30
	CURE-TSD	71.18	95.48	83.84	
	IVI-TSR	66.17	92.94	80.52	
	Overall	68.93	94.46	82.57	
SignHRNet-CA	TT-100K	69.72	95.33	83.57	29
	CURE-TSD	71.46	95.63	83.94	
	IVI-TSR	66.58	93.12	80.77	
	Overall	69.25	94.69	82.76	
SignHRNet-CBAM	TT-100K	69.05	94.71	82.95	27
	CURE-TSD	70.71	95.17	83.66	
	IVI-TSR	65.68	92.73	80.41	
	Overall	68.48	94.20	82.34	
SignHRNet-DA	TT-100K	70.81	96.06	84.11	23
	CURE-TSD	72.46	96.15	84.47	
	IVI-TSR	67.23	93.92	81.68	
	Overall	70.17	95.38	83.42	

5. Conclusion

This paper has designed an effective attentive semi-anchoring guided high-resolution network, termed as SignHRNet, for street-level traffic sign recognition tasks. The SignHRNet employed a one-stage processing architecture and consisted of an attentive HRNet as the feature extractor and a semi-anchoring guided detection head for traffic signs recognition and bounding boxes regression. To be specific, stacked with an HRNet structure boosted by a dual-attention module, the SignHRNet can export multiscale strong and task-aware feature semantics, which favored significantly the localization and identification of the traffic sign instances with diverse conditions, especially the small-size ones. Furthermore, designed with a semi-anchoring guided strategy, which used an anchor-free scheme for categorization and an anchor-based scheme for localization, the SignHRNet can achieve competitive efficiency and effectiveness with lightweight, high-quality anchors. In addition, assisted by an FAM for providing instance-sensitive feature semantics, the quality of the predicted bounding boxes was further promoted. The proposed SignHRNet has been intensively examined on three large-size datasets. Experimental results showed that an excellent overall performance with an average mAP of 72.85 %, an average mAP50 of 96.48 %, and an average mAP75 of 85.31 %, respectively, was attained for handling traffic signs of varying conditions under diverse scenarios. Ablative and comparative analyses also demonstrated the superior applicability and advanced effectiveness of the SignHRNet in traffic signs recognition tasks. However, the construction and optimization of the SignHRNet still requires large numbers of annotated data and large amount of computation resources. In our future works, we will investigate weakly supervised or few-shot strategies to well alleviate these issues.

Declaration of Competing Interest

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

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62076107, 51975239, and 41971414, in part by the Natural Science Foundation of Jiangsu Province under Grants BK20211365 and BK20191214, and in part by the Six Talent Peaks Project in Jiangsu Province under Grant XYDXX-098. The authors would like to acknowledge the anonymous reviewers for their valuable comments in improving the paper, and the volunteers for their extensive efforts in building the IVI-TSR dataset.

References

Ahmed, S., Kamal, U., Hasan, M.K., 2022. DFR-TSD: A deep learning based framework for robust traffic sign detection under challenging weather conditions. IEEE Trans. Intell. Transp. Syst. 23 (6), 5150–5162.

Almahairi, A., Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H., Courville, A., 2016. Dynamic capacity networks. arXiv preprint, arXiv:1511.07838v5. [Online]. Available: https://arxiv.org/abs/1511.07838v5.

Angles, B., Jin, Y., Kornblith, S., Tagliasacchi, A., Yi, K.M., 2021. MIST: Multiple instance spatial transformer. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 2412–2422.

Avramović, A., Sluga, D., Tabernik, D., Skočaj, D., Stojnić, V., Ilc, N., 2020. Neural-network-based traffic sign detection and recognition in high-definition images using region focusing and parallelization. IEEE Access 8, 189855–189868.

Bagi, R., Dutta, T., Nigam, N., Verma, D., Gupta, H.P., 2022. Met-MLTS: Leveraging smartphones for end-to-end spotting of multilingual oriented scene texts and traffic signs in adverse meteorological conditions. IEEE Trans. Intell. Transport. Syst. 23 (8), 12801–12810.

Boumediene, M., Lauffenburger, J., Daniel, J., Cudel, C., Ouamri, A., 2014. Multi-ROI association and tracking with belief functions: Application to traffic sign recognition. IEEE Trans. Intell. Transp. Syst. 15 (6), 2470–2479.

Cao, J., Zhang, J., Huang, W., 2021. Traffic sign detection and recognition using multi-scale fusion and prime sample attention. IEEE Access 9, 3579–3591.

Chen, C., Fan, Q., Panda, R., 2021. CrossViT: Cross-attention multi-scale vision transformer for image classification. In: Proc. Int. Conf. Comput. Vis., Virtual, pp. 357–366.

Chen, X., Yan, X., Zheng, F., Jiang, Y., Xia, S., Zhao, Y., Ji, R., 2020. One-shot adversarial attacks on visual tracking with dual attention. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 10176–10185.

Dewi, C., Chen, R., Liu, Y., Jiang, X., Hartomo, K.D., 2021. Yolo V4 for advanced traffic sign recognition with synthetic training data generated by various GAN. IEEE Access 9, 97228–97242.

Fang, H., Zhang, D., Zhang, Y., Chen, M., Li, J., Hu, Y., Cai, D., He, X., 2021. Salient object ranking with position-preserved attention. In: Proc. Int. Conf. Comput. Vis., Virtual, pp. 16331–16341.

Feng, G., Hu, Z., Zhang, L., Lu, H., 2021. Encoder fusion network with co-attention embedding for referring image segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 15506–15515.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual-attention network for scene segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 3146–3154.

Gao, M., Chen, C., Shi, J., Lai, C.S., Yang, Y., Dong, Z., 2020. A multiscale recognition method for the optimization of traffic signs using GMM and category quality focal loss. Sens. 20 (17), 4850.

Girshick, R., 2015. Fast R-CNN. In: Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, pp. 1440–1448.

González, Á., Bergasa, L.M., Yebes, J.J., 2014. Text detection and recognition on traffic panels from street-level imagery using visual appearance. IEEE Trans. Intell. Transp. Syst. 15 (1), 228–238.

Greenhalgh, J., Mirmehdi, M., 2012. Real-time detection and recognition of road traffic signs. IEEE Trans. Intell. Transp. Syst. 13 (4), 1498–1506.

Greenhalgh, J., Mirmehdi, M., 2015. Recognizing text-based traffic signs. IEEE Trans. Intell. Transp. Syst. 16 (3), 1360–1369.

Guan, H., Yu, Y., Peng, D., Zang, Y., Lu, J., Li, A., Li, J., 2020. A convolutional capsule network for traffic-sign recognition using mobile LiDAR data with digital images. IEEE Geosci. Remote Sens. Lett. 17 (6), 1067–1071.

Guo, X., Guo, X., Lu, Y., 2021. SSAN: Separable self-attention network for video representation learning. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 12618–12627.

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., Zhang, S., Martin, R.R., Cheng, M., Hu, S., 2022. Attention mechanisms in computer vision: A survey. Comput. Vis. Media 8, 331–368.

Guo, J., You, R., Huang, L., 2020. Mixed vertical-and-horizontal-text traffic sign detection and recognition for street-level scene. IEEE Access 8, 69413–69425.

He, S., Chen, L., Zhang, S., Guo, Z., Sun, P., Liu, H., Liu, H., 2021. Automatic recognition of traffic signs based on visual inspection. IEEE Access 9, 43253–43261.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 13713–13722.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2011–2023.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. In: Proc. Conf. Neural Inform. In: Process. Syst., Montreal, Canada, pp. 2017–2025.

Javanmardi, M., Song, Z., Qi, X., 2021. A fusion approach to detect traffic signs using registered color images and noisy airborne LiDAR data. Appl. Sci. 11 (1), 309.

Jin, Y., Fu, Y., Wang, W., Guo, J., Ren, C., Xiang, X., 2020. Multi-feature fusion and enhancement single shot detector for traffic sign recognition. IEEE Access 8, 38931–38940.

Kamal, U., Tonmoy, T.I., Das, S., Hasan, M.K., 2020. Automatic traffic sign detection and recognition using SegU-Net and a modified Tversky loss function with L1-constraint. IEEE Trans. Intell. Transp. Syst. 21 (4), 1467–1479.

Khan, J.F., Bhuiyan, S.M.A., Adhami, R.R., 2011. Image segmentation and shape analysis for road-sign detection. IEEE Trans. Intell. Transp. Syst. 12 (1), 83–96.

Lee, H.S., Kim, K., 2018. Simultaneous traffic sign detection and boundary estimation using convolutional neural network. IEEE Trans. Intell. Transp. Syst. 19 (5), 1652–1663.

Li, J., Wang, Z., 2019. Real-time traffic sign recognition based on efficient CNNs in the wild. IEEE Trans. Intell. Transp. Syst. 20 (3), 975–984.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollá, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327.

Liu, C., Chang, F., Chen, Z., 2014. Rapid multiclass traffic sign detection in high-resolution images. IEEE Trans. Intell. Transp. Syst. 15 (6), 2394–2403.

Liu, C., Chang, F., Chen, Z., Liu, D., 2016. Fast traffic sign recognition via high-contrast region extraction and extended sparse representation. IEEE Trans. Intell. Transp. Syst. 17 (1), 79–92.

Liu, C., Li, S., Chang, F., Wang, Y., 2019. Machine vision based traffic sign detection methods: Review, analyses and perspectives. IEEE Access 7, 86578–86596.

Liu, Y., Peng, J., Xue, J., Chen, Y., Fu, Z., 2021. TSingNet: Scale-aware and context-rich feature learning for traffic sign detection and recognition in the wild. Neurocomput. 447, 10–22.

Lu, K., Ding, Z., Ge, S., 2012. Sparse-representation-based graph embedding for traffic sign recognition. IEEE Trans. Intell. Transp. Syst. 13 (4), 1515–1524.

Luo, H., Yang, Y., Tong, B., Wu, F., Fan, B., 2018. Traffic sign recognition using a multi-task convolutional neural network. IEEE Trans. Intell. Transp. Syst. 19 (4), 1100–1111.

Mannan, A., Javed, K., Rehman, A.U., Babri, H.A., Noon, S.K., 2019. Classification of degraded traffic signs using flexible mixture model and transfer learning. IEEE Access 7, 148800–148813.

Mayo, B., Hazan, T., Tal, A., 2021. Visual navigation with spatial attention. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 16898–16907.

Meuter, M., Nunn, C., Görmer, S.M., Müller-Schneiders, S., Kummert, A., 2011. A decision fusion and reasoning module for a traffic sign recognition system. IEEE Trans. Intell. Transp. Syst. 12 (4), 1126–1134.

Møgelmose, A., Trivedi, M.M., Moeslund, T.B., 2012. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. IEEE Trans. Intell. Transp. Syst. 13 (4), 1484–1497.

Møgelmose, A., Liu, D., Trivedi, M.M., 2015. Detection of U.S. traffic signs. IEEE Trans. Intell. Transp. Syst. 16 (6), 3116–3125.

Nartey, O.T., Yang, G., Asare, S.K., Wu, J., Frempong, L.N., 2020. Robust semi-supervised traffic sign recognition via self-training and weakly-supervised learning. Sens. 20 (9), 2684.

Qin, Z., Zhang, P., Wu, F., Li, X., 2021. FcaNet: Frequency channel attention networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 783–792.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Rezatofighi, H., Tsoi, N., Gwak, J.Y., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 658–666.

Ruta, A., Li, Y., Liu, X., 2010. Robust class similarity measure for traffic sign recognition. IEEE Trans. Intell. Transp. Syst. 11 (4), 846–855.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. Int. Conf. Comput. Vis., Venice, Italy, pp. 618–626.

Serna, C.G., Ruichek, Y., 2020. Traffic signs detection and classification for European urban environments. IEEE Trans. Intell. Transp. Syst. 21 (10), 4388–4399.

Shao, F., Wang, X., Meng, F., Zhu, J., Wang, D., Dai, J., 2019. Improved faster R-CNN traffic sign detection based on a second region of interest and highly possible regions proposal network. Sens. 19 (10), 2288.

Shen, L., You, L., Peng, B., Zhang, C., 2021. Group multi-scale attention pyramid network for traffic sign detection. Neurocomput. 452, 1–14.

Song, S., Que, Z., Hou, J., Du, S., Song, Y., 2019. An efficient convolutional neural network for small traffic sign detection. J. Syst. Archit. 97, 269–277.

Sun, C., Ai, Y., Wang, S., Zhang, W., 2020. Dense-RefineDet for traffic sign detection and classification. Sens. 20 (22), 6570.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 5693–5703.

Sun, P., Zhang, W., Wang, H., Li, S., Li, X., 2021. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 1407–1417.

Tai, S., Dewi, C., Chen, R., Liu, Y., Jiang, X., Yu, H., 2020. Deep learning for traffic sign recognition based on spatial pyramid pooling with scale analysis. Appl. Sci. 10 (19), 6997.

Temel, D., Alshawi, T., Chen, M., AlRegib, G., 2019. CURE-TSD: Challenging unreal and real environments for traffic sign detection. IEEE Dataport. https://doi.org/10.21227/en9z-mq69.

Timofte, R., Zimmermann, K., Gool, L.V., 2014. Multi-view traffic sign detection, recognition, and 3D localization. Mach. Vis. Appl. 25 (3), 633–647.

Ulutan, O., Iftekhar, A.S.M., Manjunath, B.S., 2020. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 13617–13626.

Wali, S.B., Abdullah, M.A., Hannan, M.A., Hussain, A., Samad, S.A., Ker, P.J., Mansor, M. B., 2019. Vision-based traffic sign detection and recognition systems: Current trends and challenges. Sens. 19 (9), 2093.

Wang, X., Cai, Z., Gao, D., Vasconcelos, N., 2019a. Towards universal object detection by domain attention. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 7289–7298.

Wang, Z., Wang, J., Li, Y., Wang, S., 2020. Traffic sign recognition with lightweight two-stage model in complex scenes. IEEE Trans. Intell. Transp. Syst. (early access) https://doi.org/10.1109/TITS.2020.3020556.

Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C.H., Ling, H., 2019b. Learning unsupervised video object segmentation through visual attention. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 3064–3074.

Wang, L., Zhou, K., Chu, A., Wang, G., Wang, L., 2021. An improved light-weight traffic sign recognition algorithm based on YOLOv4-tiny. IEEE Access 9, 124963–124971.

Wei, L., Xu, C., Li, S., Tu, X., 2020. Traffic sign detection and recognition using novel center-point estimation and local features. IEEE Access 8, 83611–83621.

Wiles, O., Ehrhardt, S., Zisserman, A., 2021. Co-attention for conditioned image matching. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 15920–15929.

Woo, S., Park, J., Lee, J., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: Proc. European Conf. Comput. Vis., Munich, Germany, pp. 3–19.

Xie, B., Weng, X.X., 2019. Real-time embedded traffic sign recognition using efficient convolutional neural network. IEEE Access 7, 53330–53346.

Yang, Y., Luo, H., Xu, H., Wu, F., 2016. Towards real-time traffic sign detection and classification. IEEE Trans. Intell. Transp. Syst. 17 (7), 2022–2031.

Yazdan, R., Varshosaz, M., 2021. Improving traffic sign recognition results in urban areas by overcoming the impact of scale and rotation. ISPRS J. Photogramm. Remote Sens. 171, 18–35.

You, S., Bi, Q., Ji, Y., Liu, S., Feng, Y., Wu, F., 2020. Traffic sign detection method based on improved SSD. Inform. 11 (10), 475.

Yuan, X., Hao, X., Chen, H., Wei, X., 2014. Robust traffic sign recognition based on color global and local oriented edge magnitude patterns. IEEE Trans. Intell. Transp. Syst. 15 (4), 1466–1477.

Yuan, Y., Xiong, Z., Wang, Q., 2017. An incremental framework for video-based traffic sign detection, tracking, and recognition. IEEE Trans. Intell. Transp. Syst. 18 (7), 1918–1929.

Yue, L., Abdel-Aty, M.A., Wu, Y., Farid, A., 2020. The practical effectiveness of advanced driver assistance systems at different roadway facilities: System limitation, adoption, and usage. IEEE Trans. Intell. Transp. Syst. 21 (9), 3859–3870.

Zaklouta, F., Stanciulescu, B., 2012. Real-time traffic-sign recognition using tree classifiers. IEEE Trans. Intell. Transp. Syst. 13 (4), 1507–1514.

Zeng, Y., Xu, X., Shen, D., Fang, Y., Xiao, Z., 2017. Traffic sign recognition using kernel extreme learning machines with deep perceptual features. IEEE Trans. Intell. Transp. Syst. 18 (6), 1647–1653.

Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D., 2021. ESPANet: An efficient pyramid squeeze attention block on convolutional neural network. arXiv preprint, arXiv: 2105.14447v2. (Online). Available: https://arxiv.org/abs/2105.14447v2.

Zhang, J., Xie, Z., Sun, J., Zou, X., Wang, J., 2020. A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection. IEEE Access 8, 29742–29754.

Zhao, T., Wu, X., 2019. Pyramid feature attention network for saliency detection. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Long Beach, USA, pp. 3085–3094.

Zhong, Z., Lin, Z.Q., Bidart, R., Hu, X., Daya, I.B., Li, Z., Zheng, W., Li, J., Wong, A., 2020. Squeeze-and-attention networks for semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Virtual, pp. 13065–13074.

Zhou, S., Deng, C., Piao, Z., Zhan, B., 2020. Few-shot traffic sign recognition with clustering inductive bias and random neural network. Pattern Recog. 100, 107160.

Zhou, K., Zhan, Y., Fu, D., 2021. Learning region-based attention network for traffic sign recognition. Sens. 21 (3), 686.

Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S., 2016. Traffic-sign detection and classification in the wild. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog., Las Vegas, USA, pp. 2110–2118.

Zhu, Y., Liao, M., Yang, M., Liu, W., 2018. Cascaded segmentation-detection networks for text-based traffic sign detection. IEEE Trans. Intell. Transp. Syst. 19 (1), 209–219.

Zhu, K., Wu, J., 2021. Residual attention: A simple but effective method for multi-label recognition. In: Proc. Int. Conf. Comput. Vis., Virtual, pp. 184–193.