Electronic Engineering and Informatics
G. Izat Rashed (Ed.)

doi:10.3233/ATDE240100

Improved YOLOv5 with Well-Targeted Solution for Traffic Sign Detection

Guangyi WEI, Jindong XU[1], Zongbao LIANG, Qianpeng CHONG, Yijie WANG
School of Computer and Control Engineering, Yantai University, Yantai, China

Abstract. Traffic sign detection (TSD) is a significant task in the field of computer vision, which has important applications in traffic safety and driverless driving. However, this fundamental but challenging task still has always influenced by numerous negative factors, such as light intensity, severe weather and distance. In addition, this task is characterized by the small scale and irregular distribution of detected objects in complex traffic scenes. To mitigate these aforementioned challenges, this paper presents a method for TSD task improved on YOLOv5 with specific improvements. First, to alleviate the adverse impact posed by noise and enhance the feature representation, we design an adaptive network to preprocess the input image. Subsequently, we introduce recursive gated convolution and a second-order attention module in the neck to improve the sensibility of the proposed structure for small objects, i.e., traffic signs in the large-scale and complex scene. Finally, Control Distance Intersection over Union (CDIoU) loss is utilized to accelerate the convergence progress while ensuring the model detection effect. Experimental results illustrate qualitatively and quantitatively that our method outperforms other state-of-the-art detection methods in TSD.

Keywords. attention, gated convolution, traffic sign detection, YOLOv5

1. Introduction

TSD is a critical task of driverless driving and one of the key research areas in computer vision [1]. This task faces multiple challenges, with different weather conditions affecting the visibility of the signs, as well as fading, blurring and even damage to the traffic signs themselves, all of which can make detection more complex. Therefore, it is particularly important to study how to accurately detect traffic signs in complicated environments.

With the incredible explosion of CNN, many CNN-based methods were rapidly applied to the field of object detection with excellent performances. Girshick et al. presented regions with CNN feature (R-CNN) [2], which is the first two-stage object detection method based on deep learning, and this method significantly outperforms the traditional methods. Subsequent object detection methods, such as single shot multi-box detector (SSD) [3], Faster R-CNN [4] and you only look once (YOLO) series [5][6][7], obtained better effect in object localization and classification tasks. Cui et al. [8] presented CAB-s Net for TSD. However, these architectures generally aim to extract more hard-to-dig features by building deeper network structures, with large models, slow detection speeds, and the need for a large amount of hardware resource support, making

[1] Corresponding author: Jindong XU, School of Computer and Control Engineering, Yantai University, e-mail: xujindong1980@163.com

it difficult to be applied in mobile devices.

Aiming at the challenges of the TSD task, this paper presents a new TSD method improved on YOLOv5, which improves the detection accuracy and reduces the leakage rate. All in all, the main contributions of this paper are as follows:

a) We summarize the difficulties and challenges in the field of TSD and make a specific detection strategy.

b) A lightweight preprocessing network is utilized to reduce the noise of input images. A contextual semantic module, and a second-order attention mechanism are introduced into neck, which greatly improves the level of attention and perceptual localization of traffic signs. We introduce the loss function CDIoU to accelerate convergence process.

c) Experiments show that our method can better perceive and detect the more difficult samples in the TSD task, proving the superiority of our method.

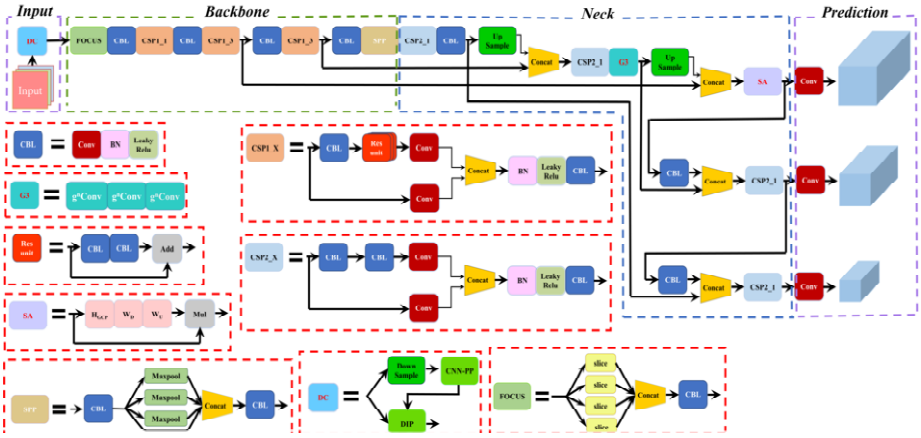

Figure 1. The overall network architecture

2. Methodology

2.1 Overview

The workflow of the proposed architecture can be seen in Figure 1. Firstly, original images are pre-processed by DC module to diminish the difficulty of extracting the feature representation of the image by the backbone. In the neck, the contextual semantic module G3 is introduced to provide a more comprehensive and accurate semantic representation to enhance the understanding and localization capability of the network. A second-order attention SA is introduced into neck get more accurately focus on and capture the important features of the target region, targeting the detection capability and finally outputting the detection results. In addition, the CDIoU loss function is used to speed up model convergence.

2.2 Adaptive Noise Alleviation Module DC

It is difficult to detect objects from low-quality images under bad weather conditions,

and to alleviate this problem, we propose the differentiable image processing combined with a small convolutional neural network module (DC) for adaptive noise reduction of the input image. DC is inspired by the literature [9], utilizing a small discriminative network, which dynamically adapts to the noise distribution in the image, making it clearer and reducing the noise factors in the image. DC consists of a differentiable image processing (DIP) module plus a tiny convolutional neural network (CNN-PP). Since CNN-PP optimizes for gradients, the filters of the images need to be differentiable to allow training the network by back-propagation. Convolutional neural networks consume a lot of computational resources when processing highly differentiated images, CNN-PP downsamples the images to learn the parameters of the filters and later passes these parameters back to the DIP module to be applied to the original resolution images.

2.3 Contextual Semantics Module G3

The contextual semantic module G3 consists of three recursively gated convolutions in series, with the output of the former convolution being the input to the latter. Recursive gated convolution [10] is composed of convolution operation, linear projection and element-by-element multiplication and can be used to implement long-term and higher-order spatial interactions with similar self-attention for adaptive spatial mixing of inputs. This function utilizes the first-order spatial interaction of the gated convolution. Let $x \in \mathbb{R}^{HW\times C}$ be the input factor and the output $y = gConv(x)$ can be written as:

$$[\boldsymbol{m}_0^{HW\times C}, \boldsymbol{n}_0^{HW\times C}] = \phi_{in}(x) \in \mathbb{R}^{HW\times 2C}$$

$$\boldsymbol{M}_1 = f(\boldsymbol{n}_0)\odot \boldsymbol{m}_0 \in \mathbb{R}^{HWC}, y = \phi_{out}(\boldsymbol{m}_1) \in \mathbb{R}^{HW\times C} \quad (1)$$

where ϕ_{in}, ϕ_{out} are linear projection that achieve channel mixing and are convolutional layers in the depth direction. Note that $m_1^{(i,c)} = \sum_{j\in \psi i} \mathcal{W}_{i\to j}^{c} n_0^{(j,c)} m_0^{(i,c)}$ where ψ_i is the local window centered on i and $\mathcal{W}$ denotes the convolution weight of f. Hence, Eq. (1) explicitly explains the interaction between contiguous factors $m_0^{(i)}$ and $n_0^{(i)}$ by element-by-element multiplication, which can better model more complex spatial interactions. Higher order interactions are then introduced. Projection features m_0 and $\{n_k\}_{k=0}^{n-1}$: is obtained using ϕ_{in}, as shown in Eq. (2).

$$\left[\boldsymbol{m}_0^{HW\times C_0}, \boldsymbol{n}_0^{HW\times C_0}, \dots, \boldsymbol{n}_0^{HW\times C_{n-1}}\right] = \phi_{in}(x) \in \mathbb{R}^{HW\times(\sum_{0\le k\le n-1} C_k)} \quad (2)$$

After that, the gated convolution is performed by Eq. (3):

$$\boldsymbol{m}_{k+1} = f_k(\boldsymbol{n}_k)\odot g_k(\boldsymbol{m}_k)/\alpha, k = 0,1,\dots,n-1 \quad (3)$$

where the output is multiplied by a scaling factor $1/\alpha$ as a way to stabilize the training.

$$g_k = \begin{cases} Identity, k = 0, \\ Linear(C_{k-1}, C_k), 1 \le k \le n-1. \end{cases} \quad (4)$$

Finally, the output of n_n to the projection layer ϕ_{out} to obtain $\mathrm{g_n Conv}$'s input is the feature map with channel C , after the first convolutional layer, the number of channels becomes twice as many, and the output of the first convolution is split into two sections, the first is used for the next layer, the second is sent to the output of the depthwise separable convolution, and the output of the depthwise separable convolution is used as the input of the remaining three layers. As in Eq. (4), g_k are utilize to match the dimension in orders.

2.4 Second-order Attention SA

Traffic signs are easy to be ignored during the feature extraction process, inspired by [11], we introduce the second-order attention (SA) into neck. Firstly, the covariance matrix is calculated to represent the correlation between different channels. Based on the mean and variance of each channel, the adaptive scale factor is calculated. The scale factor is used to weight the feature matrix at the element level to enhance the important channel features. SA can adaptively realign the features exploiting higher-than-first-order feature statistics to enhance the learning ability of detection, provide better perception of small objects, and enhance the detection performance.

2.5 Loss Function $\mathcal{L}_{CDIoU}$

As a result of environmental influences, the spatial configuration and morphology of objects within the input images can be subject to modification. YOLOv5 utilizes the CIoU loss function, which incorporates the center distance of the bounding box, discrepancies in width and height, as well as overlapping areas. However, utilizing aspect ratio as an influencing factor causes the regression results of the loss function to deviate from the correct regression objective. Considering the shortcomings of CIoU, we introduced the loss function CDIoU [12], which uses a new evaluation method to calculate the loss between GT and RP, which significantly improves the computational efficiency and accuracy without increasing the computation time, and is more suitable for the current task. The calculation method is shown in Eqs. (5-7) below:

$$diou = \frac{\|RP-GT\|_2}{4MBR's\ diagonal} = \frac{AE+BF+CG+DH}{4WY}, \tag{5}$$

$$CDIoU = IoU + \lambda(1 - diou) \tag{6}$$

$$\mathcal{L}_{CDIOU} = \mathcal{L}_{IoU} + diou \tag{7}$$

where *MBR* is the smallest external rectangle enclosing the two boxes, *WY* denotes the diagonal of *MRB*, *ABCD* and *EFGH* represent the four vertices of GT and RP, respectively, and $\mathcal{L}_{CDIoU}$ is the loss function based on CDIoU. Therefore, the higher value of CDIoU, the higher similarity. In the process of weight iteration, the model continuously pulls the four vertices of RP toward the four vertices of GT until they overlap.

3. Experiments

3.1 Dataset and Evaluation Metrics

To evaluate our method, we selected the Chinese TSD dataset CCTSDB [13] and used four metrics: Precision, Recall, mAP_0.5, and mAP_0.5:0.95, to evaluate the detection results. Chinese traffic signs are divided into three categories in CCTSDB: blue mandatory signs, yellow danger signs, and red prohibited signs. In this experiment, we selected some images with distinctive scene features in the CCTSDB as the training set and verified the performance of the method on the test set.

Table 1. Evaluation results for each metric on the CCTSDB dataset comparing our method with advanced target detection methods. The bold format indicates the best results for the network in the evaluation metrics.

Method	Precision (%)	Recall (%)	mAP_0.5 (%)	mAP_0.5:0.95 (%)
Faster R-CNN	74.12	91.95	92.61	**76.21**
SSD	70.41	86.15	87.21	72.31
YOLOv3	83.21	78.50	83.47	61.45
YOLOv4	90.84	91.81	92.73	70.27
YOLOv5-s	93.16	95.21	96.96	73.12
ours	**93.22**	**96.38**	**97.27**	72.56

3.2 Comparison Experiments

We utilize several advanced object detection method to compare with us, including Faster R-CNN[4], SSD[3], YOLOv3[6], YOLOv4[7], and YOLOv5-s. The results in Table 1 demonstrate that our method outperforms the detection performance on the CCTSDB dataset.

The object detection performance of our method and YOLOv5-s on the Chinese Traffic Sign Detection and Benchmark (CCTSDB) dataset is shown in Figure 1. These key objects are small in size, variable in location, and heavily disturbed by background noise. Compared with YOLOv5-s, our proposed method has more advantages and the results are satisfactory.

Figure 2. Results of our method on the CCTSDB dataset compared with YOLOv5, the red boxes mark the traffic signs missed by YOLOv5

4. Conclusions

In this paper, we discuss the difficulties of TSD, propose targeted solutions and make improvements based on YOLOv5. Firstly, the original input image is input to the detection network after weakening the noise. Secondly, the contextual semantic module G3 and the second-order attention mechanism SA are introduced into neck to enhance the understanding and localization ability of the network to strengthen the attention to traffic signs. In addition, $\mathcal{L}_{CDIoU}$ is introduced to boost convergence process of our method. The experimental results show the high level achieved by our proposed method in detecting traffic signs in complex environments. In future research, we try to add a priori knowledge from real scenarios of TSD to the method to assist detection and further improve the robustness and generalization ability of the method.

References

[1] Radu, M.D., Costea, I.M., Stan, V.A. (2020) Automatic Traffic Sign Recognition Artificial Intelligence—Deep Learning Algorithm. In: Proceedings of the 2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). Bucharest, Romania. pp. 1–4. https://doi.org/10.1109/ECAI50035.2020.9223186

[2] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA. pp. 580–587. https://doi.org/10.48550/arXiv.1311.2524.

[3] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016) SSD: Single Shot Multibox Detector. In: Proceedings of the European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands. pp. 21–37. https://doi.org/10.1109/TPAMI.2016.2577031.

[4] Ren, S., He, K., Girshick, R., Sun, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems. New York. Volume 28. https://doi.org/10.1109/TPAMI.2016.2577031.

[5] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas. pp. 779–788. https://doi.org/10.48550/arXiv.1506.02640.

[6] Redmon, J., & Farhadi, A. (2018) YOLOv3: An incremental improvement. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Salt Lake City. pp. 7263-7271.
https://doi.org/10.48550/arXiv.1804.02767.

[7] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Seattle. pp. 1563-1572. https://doi.org/10.48550/arXiv.2004.10934.

[8] Cui, L., Lv, P., Jiang, X., Gao, Z., Zhou, B., Zhang, L., Shao, L., Xu, M. (2022) Context-Aware Block Net for Small Object Detection. IEEE Transactions on Cybernetics, vol. 52, no. 4, pp. 2300-2313. https://doi.org/10.1109/TCYB.2020.3004636.

[9] Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., & Zhang, L. (2021) Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. AAAI Conference on Artificial Intelligence.
https://doi.org/10.48550/arXiv.2112.08088.

[10] Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S.N., & Lu, J. (2022) HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. ArXiv, abs/2207.14284. https://doi.org/10.48550/arXiv.2207.14284.

[11] Dai, T., Cai J.R, Zhang, Y.B, Xia, S.T., Zhang, L. (2019) Second-Order Attention Network for Single Image Super-Resolution. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach. pp. 11057-11066, DOI: 10.1109/CVPR.2019.01132.

[12] Chen, D., & Miao, D. (2021) Control Distance IoU and Control Distance IoU Loss Function for Better Bounding Box Regression. ArXiv, abs/2103.11696. https://doi.org/10.48550/arXiv.2103.11696.

[13] Zhang, J.M., Zou, X., Kuang, L.D, Wang, J., Sherratt, R.S., Yu, X.F. (2022) CCTSDB 2021: A more comprehensive traffic sign detection benchmark. Human-centric Computing and Information Sciences. vol.12, pp. 23. DOI: 10.22967/HCIS.2022.12.023.