Multi-scale detector optimized for small target*

ZHU Yongchang[1], YANG Sen[1], TONG Jigang[1], and WANG Zenghui[2,3]**

1. School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin 300384, China

2. Center of Artificial Intelligence and Data Science, University of South Africa, Florida 1709, South Africa

3. Department of Electrical Engineering, University of South Africa, Florida 1709, South Africa

(Received 9 July 2023; Revised 18 December 2023)

The effectiveness of deep learning networks in detecting small objects is limited, thereby posing challenges in addressing practical object detection tasks. In this research, we propose a small object detection model that operates at multiple scales. The model incorporates a multi-level bidirectional pyramid structure, which integrates deep and shallow networks to simultaneously preserve intricate local details and augment global features. Moreover, a dedicated multi-scale detection head is integrated into the model, specifically designed to capture crucial information pertaining to small objects. Through comprehensive experimentation, we have achieved promising results, wherein our proposed model exhibits a mean average precision (*mAP*) that surpasses that of the well-established you only look once version 7 (YOLOv7) model by 1.1%. These findings validate the improved performance of our model in both conventional and small object detection scenarios.

Document code: A **Article ID:** 1673-1905(2024)04-0243-6

DOI https://doi.org/10.1007/s11801-024-3126-1

Deep learning-based object detection constitutes a fundamental task in the domain of computer vision, with the objective of precisely recognizing and localizing specific objects within images or videos. Leveraging deep neural network models, such algorithms automatically extract discriminative features from input visual data and subsequently detect the objects of interest by virtue of acquiring substantial knowledge from extensive annotated training datasets[1-3].

Encountering small objects during the process of deep learning object detection engenders formidable challenges[4,5], effectively impeding the attainment of superior precision and inevitably impacting the model's inference performance. Notable hurdles governing this issue encompass the following aspects[6-8].

Size issue: Small objects generally possess diminutive spatial dimensions, often measuring merely a few pixels in comparison to the overall image. Consequently, their detection and localization become arduous as intricate details and characteristic features may become either obscured or lost within the image.

Feature representation: Small objects exhibit features that are typically fainter or more ambiguous in nature, leading to less precise or comprehensive representations acquired by deep learning models. Traditional convolutional neural networks tend to exhibit a bias towards accommodating larger objects in their design, rendering them less adept at extracting features from small-sized objects.

Limited semantic information: Small objects frequently present limited availability of informative semantic cues. Owing to their restricted levels of details, the features of small objects may become entangled with the surrounding background, thereby impeding their discernibility. This insufficiency of discriminative information consequently hinders the model's ability to effectively recognize and classify small objects.

Class imbalance issue: Within the realm of object detection, small objects are often less prevalent compared to their larger counterparts. This class imbalance can lead to an overemphasis on larger objects during the training process, consequently marginalizing the attention devoted to small objects and compromising their detection capabilities.

Consequently, the detection of small objects is notably susceptible to issues such as missed detection and false positives[9-12]. To enhance accuracy while maintaining computational efficiency, a multi-level bidirectional pyramid structure is used. Our algorithm holds the potential to enhance accuracy while maintaining computational efficiency, demonstrating superior performance in small target detection. Through effectively reducing rates of missed detection and false positives, it affords notable improvements not only in detecting objects of conventional sizes but also in the context of detecting small targets, offering significant advantages for such tasks.

* This work has been partially supported by the National Natural Science Foundation of China (No.62103298), and the South African National Research Foundation (Nos.132797 and 137951).

** E-mail: tjgtjut@163.com

The main contributions of this paper are as follows.

Deployed multi-scale detectors: The detector and its adjoint structure are reconstructed, creating a network with a different structure from the traditional three-detector network, which can improve detection capabilities at different scales.

A novel multi-scale feature fusion bi-directional feature pyramid is proposed in this study. The bi-directional feature pyramid network is reconstructed to facilitate the fusion of multi-scale features, enabling the transfer of local detailed feature information while maximizing the retention of global feature information. To enhance the feature fusion capability across multiple scales, the network incorporates jump connections.

An optimized spatial pyramidal pooling structure is introduced in this study. The utilization of a fast spatial pyramid pooling method aims to decrease computational complexity while maintaining computational accuracy, thereby enhancing the computational efficiency of the model.

Parameter calculation volume: By comparing the analysis of the amount of model computing parameters of different structures, it can reflect the optimization in the amount of network computation, while achieving the purpose of improving the speed of network computation.

Deep learning-based object detection networks refer to network models that are built upon deep learning techniques and used to detect and localize specific objects or targets in images or videos.

You only look once (YOLO)[13] treats object detection as a regression problem. It achieves object detection by dividing the entire image into grids and predicting bounding boxes and class probabilities within each grid cell. YOLO series include YOLO9000[14], YOLO version 3 (YOLOv3)[15], and YOLO version 4 (YOLOv4)[16], among others. With each version release, the YOLO series continually improves accuracy and performance by incorporating more powerful backbone networks such as DarkNet, DarkNet-53, and CSPDarkNet, as well as introducing multi-scale predictions, cross-feature layer connections, and attention mechanisms.

Neural architecture search-feature pyramid network (NAS-FPN)[17] series leverage neural architecture search (NAS) techniques to automatically search for the optimal backbone network architecture for object detection. It combines the searched architecture with the feature pyramid network (FPN) to enhance the performance of object detection. NAS-FPN further improves the performance and accuracy of object detectors by employing search algorithms to optimize network depth, feature map size, and channel numbers. Furthermore, NAS-FPN introduces the bi-directional feature pyramid network (BiFPN) structure to strengthen the feature pyramid network.

EfficientDet[18] is a series of object detectors proposed by the Google team in 2020. It utilizes EfficientNet as the backbone network and combines BiFPN with EfficientHead to achieve efficient and accurate object detection. EfficientDet series achieved better object detection performance by leveraging the efficient feature extraction capability of EfficientNet and the multi-scale feature fusion of BiFPN. The different versions of the series, ranging from EfficientDet-D0 to EfficientDet-D7, strike a balance between network scale and accuracy, making them suitable for various application needs.

These series have played a pivotal role in the advancement of object detection methods in the field, propelling detection algorithms towards higher levels of performance and efficiency.

Many researchers have conducted extensive studies to address the challenges posed by the low resolution and vulnerability to noise of small targets. In Ref.[19], an adaptive anchor box structure was proposed, which enhanced the learning capability of the model while achieving cost reduction effects. Similarly, in Ref.[20], an oversampling technique employing the copy method was introduced to enhance the detection performance for small targets, particularly when dealing with datasets that have limited samples containing such targets.

Improving the detection of small targets necessitates optimizing not only data processing techniques but also the underlying model structure. To this end, Ref.[21] presented an approach that tackled small target detection in urban scenes through joint optimization of data processing methods and model structures. Meanwhile, Ref.[22] adopted a functionally enhanced network trained in a self-supervised manner to counteract the inherent challenge of relatively low signal-to-noise ratio experienced by small targets. Additionally, Ref.[23] proposed a feature pyramid fusion network based on attention mechanism, which effectively preserved key information of small targets and proved effective for addressing the task of detecting small targets.

In this study, we optimize the model structure by incorporating a multi-scale detection head and a path aggregation network. These enhancements aim to maximize the preservation of critical information pertaining to small targets, consequently improving their detectability.

The input image is first processed by the backbone network to extract feature information from different levels of abstraction. The extracted features are then fed into the neck network, which combines and fuses them to obtain a more comprehensive representation of the object. Finally, the prediction head generates the bounding boxes and class probabilities based on the fused features and output the final results. This paper proposes a new optimized network structure, which can be divided into four parts, input, backbone, neck, and prediction as shown in Fig.1.

The input component of the proposed algorithm applies data augmentation techniques such as flipping, filling, and Mosaic splicing to enrich the dataset samples. Then, the original input samples are computed to obtain the initial prior frame and compared with ground-truth to calculate the difference and perform the reverse update.

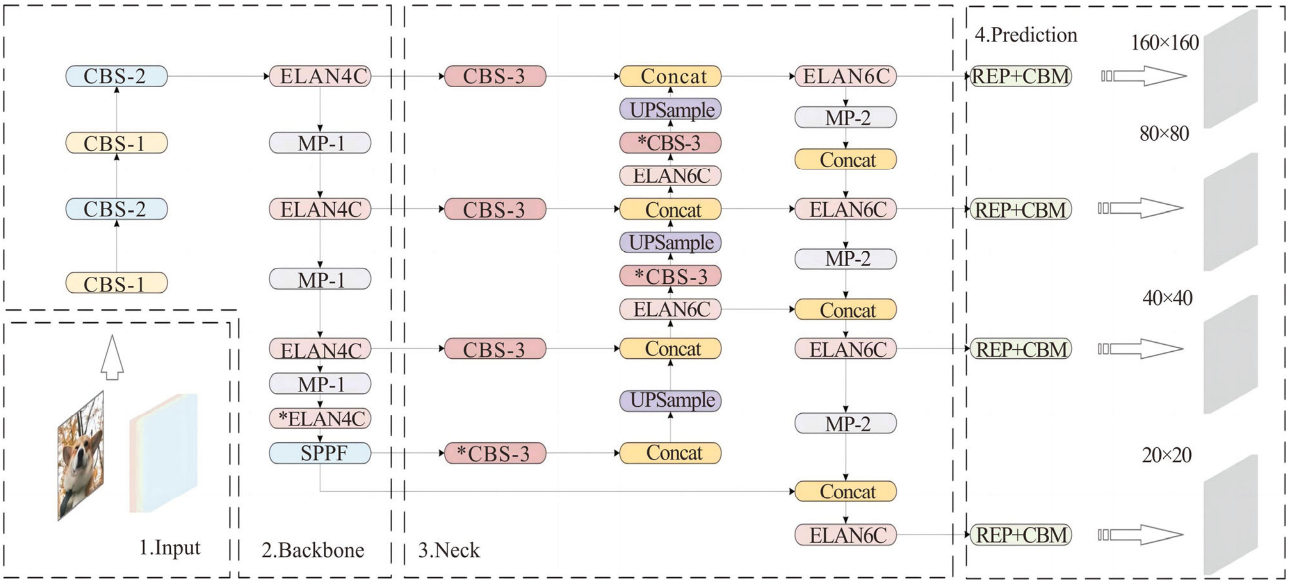

Fig.1 EW network model structure

The backbone component of our proposed framework primarily comprises several key modules, namely CBS convolution blocks, down sampling modules, feature stacking modules, and fast spatial pyramid pooling (SPPF) modules. These modules collectively play crucial roles in facilitating effective feature extraction, extracting essential information, and suppressing noise within the original input samples.

Research has indicated that incorporating shorter connections between layers proximate to the input and those close to the output of the network can render the network considerably deeper, more precise. As shown in Fig.2, efficient layer aggregation network (ELAN) module can provide the model with rich gradient information while ensuring the model parameters are lightweight.

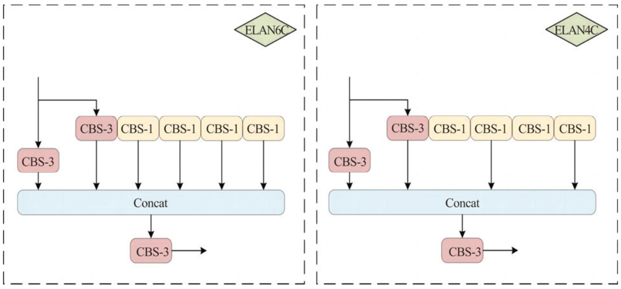

Fig.2 ELAN structure

To reduce the amount of model computation, the SPPF structure is applied, as shown in Fig.3. It can improve the computing speed of the network to some extent.

In order to better retain the semantic information of the shallow network is to maximize the reduction of the distortion rate of the model in the operation, so as to achieve the purpose of reducing the rate of missed detection and false detection, as shown in Fig.4.

In neural networks, after undergoing layer-by-layer down sampling, input images will lose some features. However, deep networks that have undergone multiple down sampling processes can better reflect global features. Shallow networks that have not undergone down sampling or have undergone fewer down sampling processes are not likely to cause the loss of key feature information for small-sized targets. They can retain local detailed features, which are important means for recognizing small targets. Therefore, shallow networks are crucial for identifying small-sized targets.

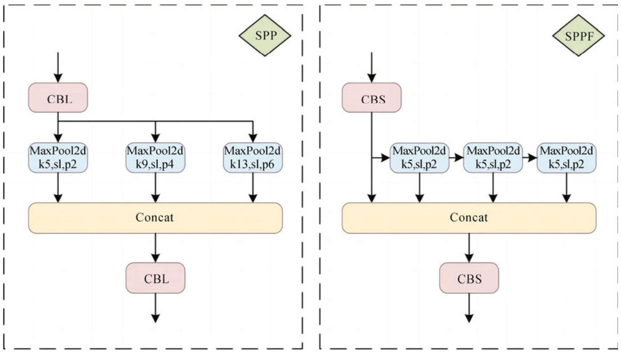

Fig.3 SPP structure and SPPF structure

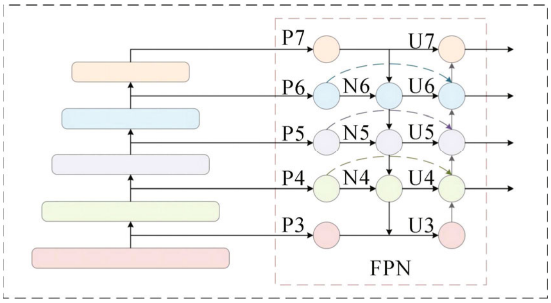

Fig.4 Path-aggregation network structure

As shown in Fig.5, in conventional neural network models, fusion features from layers P5 to P7 are used, neglecting the importance of shallow networks. In our algorithm proposed in this paper, fusion is conducted from layers P4 to P7, which enables the extraction of information from shallow networks to improve the model's sensitivity to small-sized targets.

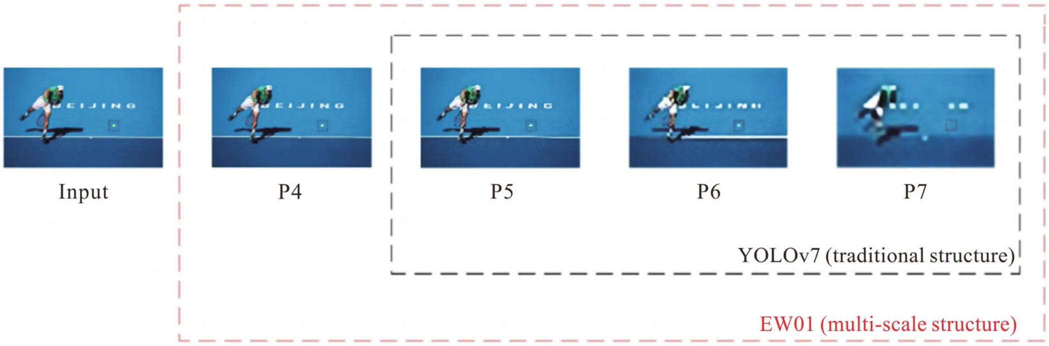

Fig.5 Sampling feature maps at various levels of the network

Furthermore, we have tried different schemes such as P3 to P6 layers and P3 to P7 layers. The actual results have shown that the fusion of P4 to P7 layers is the optimal solution in terms of comprehensive accuracy and model parameter considerations.

Adding an additional detection head at 1/4 scale allows the network to better detect smaller objects that can result in an overall improvement in the mean average precision (*mAP*) index.

The prediction component uses GIoU_Loss as the loss function of the bounding box as follows

$$L_{GIoU} = IoU + \frac{\left|C - (B^{GT} \cup B)\right|}{\left|C\right|}, \qquad (1)$$

where *IoU* denotes the intersection ratio between the prediction frame and the real frame, B^{GT} and B denote the sizes of the real frame and the predicted frame, and C is the minimum convex set enclosing B^{GT} and B.

The experiment was conducted in the framework of Pytorch, a 64-bit Windows 10 operating system, with the software and hardware platform parameters shown in Tab.1.

Tab.1 Experimental environment

Configuration	Version
GPU	NVIDIA GeForce RTX A5000
CPU	Intel(R) Xeon(R) Platinum 8358P CPU@
CUDA	12.0
Python	1.11.0

The epoch is 300, the batch size is 64, and the image size is 640×640.

The PASCAL VOC dataset was released in 2015, and the two versions of VOC2007 and VOC2012 are more frequently used in academia. It is a dataset mainly used for image classification, target detection, and image segmentation. The number of target categories in this dataset is all 20, mostly in real scenes, such as people, bicycles, cats, dogs, etc. This dataset has good image quality and complete labels, and is mostly used for model performance evaluation.

The Microsoft COCO (MS COCO) dataset is a comprehensive and extensively utilized dataset that finds widespread applications in image detection, semantic segmentation, and image captioning within the field of computer science and artificial intelligence. This dataset encompasses a substantial collection of over 330k images, out of which 220k images have been meticulously annotated. These annotations cover a staggering 1.5 million targets, which include 80 distinct object categories, such as pedestrian, car, and elephant, as well as 91 stuff categories including grass, wall, and sky. The sheer scale and diversity of this dataset make it an invaluable resource for researchers and practitioners seeking to advance their knowledge and capabilities in computer vision tasks.

The traffic camera object detection (TCOD) dataset belongs to the small target dataset, and the dataset contains only one category of car, including 6 121 images, which can well detect the sensitivity of the model to small targets, and this paper also selects this dataset for training and detection experiments.

In this research, the MS COCO dataset was chosen to examine the performance of the model on conventional targets, and the TCOD dataset was used to examine the effectiveness of the model in detecting small targets.

Tab.2 Dataset

Dataset	Year	Number	Cases	Image size	Note
PASCAL VOC 2007	2010	9 963	20	500×375	Open
PASCAL VOC 2012	2015	11 540	20	470×380	Open
MS COCO	2014	328 000	91	640×480	Open
TCOD	2019	6 121	1	416×416	Open

In order to evaluate the performance of diverse object detection algorithms, it is crucial to employ several assessment metrics, including but not limited to *mAP*, average precision (*AP*), precision (*P*), and recall (*R*).

Precision refers to the ratio of accurate positive predictions to all positive predictions conducted by the model. A higher precision value indicates a lower number of false positives, whereas a lower precision value implies a higher ratio of false positives.

Recall illustrates the percentage of truly predicted positive samples among all actual positive samples, with a higher recall value indicating fewer false negatives, thus implying that fewer positive samples are omitted. Conversely, a lower recall value denotes a higher rate of false negatives, indicating that more positive samples are missed.

In order to evaluate the efficiency of the optimized multi-scale detector algorithm (EW) for small targets detection, this study utilizes several evaluation metrics, including *mAP*, *P*, and *R*, alongside model size as an additional assessment criterion. The experimental approach involves a comparison and ablation evaluation that gauges the effectiveness of the algorithm.

According to the statistics of the MS COCO dataset, small objects have a relatively high proportion in object detection tasks. Depending on different definitions and thresholds, small objects can account for 30% to 50% or even more of the total number of objects. This depends on the standards and definitions used to determine what size of objects are considered "small". Therefore, small objects are an important subset within the MS COCO dataset and pose certain challenges to the accuracy and robustness of object detection algorithms.

Based on the presented Tab.3, it can be inferred that the proposed model achieves a 1.1% improvement in *mAP* value compared to the YOLOv7 model by employing a richer semantic overlay, and by organically integrating the deeper and shallower networks to control global and detailed features. Moreover, the proposed model surpasses other classical network models, which highlights its effectiveness in classical detection tasks.

Tab.3 Comparison of real-time object detectors (MS COCO)

Method	Param	FLOPs	AP^{val}	AP_{50}
YOLOv5-M (r6.1)	21.2M	49.0G	45.4%	-
YOLOv5-L (r6.1)	46.5M	109.1G	49.0%	-
PPYOLOE-M	23.4M	49.9G	48.6%	66.5%
PPYOLOE-L	52.2M	110.1G	50.9%	68.9%
YOLOX-M	25.3M	73.8G	46.9%	-
YOLOX-L	54.2M	155.6G	49.7%	-
YOLOv7	**36.9**M	104.7G	(**48.1%**)	(**65.4%**)
EW01	**37.4**M	119.7G	(**48.8%**)	(**66.5%**)

The data in parentheses in the table are the results of our run, and the results in the table indicate that the *mAP* value of the algorithm is improved by 1.1% compared to the YOLOv7 model.

Based on the presented Tab.4, compare our model with YOLOv7 and other publicly available models optimized for small targets. Our model achieves better performance in the Pascal VOC dataset. Additionally, comparing against other models in the TCOD dataset with numerous small targets, our model demonstrates higher accuracy in comparison to other small-sized target models.

Tab.4 Comparison with other small-scale target detection networks (Pascal VOC & TCOD)

Method	*mAP* (%)		
	VOC 2007	VOC 2012	TCOD
NAS-FPN R-50 (7@256) 640	80.5	80.1	85.2
NAS-FPN R-50 (7@256) 1 024	83.7	83.4	86.7
EifficientDet-D5+AA	85.1	84.2	84.7
EifficientDet-D6+AA	85.9	85.2	85.4
YOLOv7	85.4	84.7	83.0
EW01	**87.0**	**86.1**	**89.2**

When detecting the same content, a comparative analysis between the proposed algorithm and the YOLOv7 model yielded the following results.

Small target detection analysis: The proposed algorithm maintains high detection accuracy for larger targets while outperforming the YOLOv7 model in detecting distant and small targets.

Missing object detection analysis: The proposed algorithm significantly reduces the likelihood of missed detections due to imbalances between foreground and background categories. Moreover, the algorithm exhibits advantages in target localization and recognition classification compared to the YOLOv7 model.

Dense small target detection analysis: When compared to the YOLOv7 model, the proposed algorithm leverages deeper and shallower network information to extract more detailed global features, resulting in better performance on dense targets.

According to Tab.5, it can be inferred that the addition of multiple detection heads alone, without the incorporation of a multi-scale network structure, does not result in a substantial improvement. As such, it is recommended that multi-scale detection heads and their associated structures can be implemented concurrently within the network.

Tab.5 The results of ablation experiments

Multi-scale detection head	Multi-scale network structure	SPPF	Param	AP_{50}
√			37 842 046	65.3%
	√		38 072 424	66.2%
		√	37 622 682	65.4%
√	√		38 204 824	66.5%
	√	√	38 072 424	66.2%
√		√	37 842 046	65.3%
√	√	√	38 204 824	66.5%

In order to address the issue of low accuracy in detecting small objects in object detection tasks, a structure that integrates a multi-scale feature network is proposed to mitigate the loss of feature information caused by down sampling small objects in the model. By enhancing the influence of the shallow network and incorporating multi-scale detection heads for analyzing larger-sized feature maps, the network demonstrates effective perception of small objects, thereby improving the model's sensitivity towards small objects and achieving the goal of enhancing model accuracy. In the future, we will work on the correction of the loss function of small targets, so that the conventional network model can give better applicability to small targets, while optimizing the network structure to improve the speed.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] LIAO Y R, WANG H N, LIN C B, et al. Research progress of optical remote sensing image target detection based on deep learning[J]. Journal on communications, 2022, 43(5): 190-203.

[2] ZHANG T, LI Z, SUN Z, et al. A fully convolutional anchor-free object detector[J]. The visual computer, 2023, 39(2): 569-580.

[3] MOHAMMADKARIMI M, MEHRABI M, ARDAKANI M, et al. Deep learning-based sphere decoding[J]. IEEE transactions on wireless communications, 2019, 18(9): 4368-4378.

[4] LI Z, GUO Q, SUN B, et al. Small object detection methods in complex background: an overview[J]. International journal of pattern recognition and artificial intelligence, 2023, 37(2): 2350002.

[5] LI R, HU J, LI S, et al. Blind detection of communication signals based on improved YOLO3[C]//2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), April 9-11, 2021, Xi'an, China. New York: IEEE, 2021: 424-429.

[6] VARA P N R S, D'SOUZAKEVIN B, BHARGAVAVIJAY K. A downscaled faster-RCNN framework for signal detection and time-frequency localization in wideband RF systems[J]. IEEE transactions on wireless communications, 2020, 19(7): 4847-4862.

[7] WAN Y, LIAO Z, LIU J, et al. Small object detection leveraging density-aware scale adaptation[J]. The photogrammetric record, 2023, 38(182): 160-175.

[8] QIN H, WU Y, DONG F, et al. Dense sampling and detail enhancement network: improved small object detection based on dense sampling and detail enhancement[J]. IET computer vision, 2022, 16(4): 307-316.

[9] XIAO Z H, DONG E Z, TONG J G, et al. Light weight object detector based on composite attention residual network and boundary location loss[J]. Neurocomputing, 2022, 494: 132-147.

[10] ZHANG S F, WANG Q, ZHU T, et al. Detection and classification of small traffic signs based on cascade network[J]. Chinese journal of electronics, 2021, 30(4): 727-735.

[11] CHEN S, LI Z, TANG Z. Relation R-CNN: a graph based relation-aware network for object detection[J]. IEEE signal processing letters, 2020, 27: 1680-1684.

[12] XU D, GUAN J, FENG P, et al. Association loss for visual object detection[J]. IEEE signal processing letters, 2020, 27: 1435-1439.

[13] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[14] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 7263-7271.

[15] REDMON J, FARHADI A. YOLOV3: an incremental improvement[EB/OL]. (2018-04-08) [2023-09-05]. https://arxiv.org/abs/1804.02767.

[16] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOV4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23) [2023-09-05]. https://arxiv.org/abs/2004.10934.

[17] JIN H, SONG Q, HU X. Auto-Keras: efficient neural architecture search with network morphism[EB/OL]. (2018-06-27) [2023-09-05]. https://arxiv.org/abs/1806.10282v2.

[18] TAN M, PANG R, LE Q V. EfficientDet: scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 10778-10787.

[19] LI F, GAO D, YANG Y, et al. Small target deep convolution recognition algorithm based on improved YOLOv4[J]. International journal of machine learning and cybernetics, 2023, 14(2): 387-394.

[20] BOSQUET B, CORES D, SEIDENARI L, et al. A full data augmentation pipeline for small object detection based on generative adversarial networks[J]. Pattern recognition: the journal of the pattern recognition society, 2023, 133: 108998-109010.

[21] YANG Z, YU H, FENG M, et al. Small object augmentation of urban scenes for real-time semantic segmentation[J]. IEEE transactions on image processing, 2020, 29: 5175-5190.

[22] LEE G, HONG S, CHO D. Self-supervised feature enhancement networks for small object detection in noisy images[J]. IEEE signal processing letters, 2021, 28: 1026-1030.

[23] ZHANG H, DU Q, QI Q, et al. A recursive attention-enhanced bidirectional feature pyramid network for small object detection[J]. Multimedia tools and applications, 2023, 82(9): 13999-14018.