Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 2, 2023

Review of Small Target Detection based on Deep Learning

Heng Zhang [1, *], Wei Fu [2], Ke Wu [3]

[1] School of Computer and Software Engineering, Xihua University, Chengdu, Sichuan, 610039, China
[2] College of Chinese & Asean Arts, Chengdu University, Chengdu, Sichuan, 610106, China
[3] Harbin College Teacher Education College, Harbin University, Harbin, Heilongjiang, 150080, China

*** Corresponding author:** Heng Zhang (Email: hengzhang_xhu@163.com)

Abstract: With a large number of applications of target detection in daily life, the performance requirements of target detection are constantly improving. Many challenges about target detection have been put forward constantly, such as imbalanced samples, fewer pixels, and occlusion of the detected target, all of which bring difficulties for the model to correctly identify the target. Small target detection has always been a difficult point and research hotspot. In recent years, many algorithms for small target detection have been proposed, such as data enhancement, feature fusion, attention mechanism, and super-resolution network structure. According to the characteristics of different network structures, the training strategy can be appropriately adjusted and then applied to different environments, which can greatly improve the detection accuracy of small targets. This paper will introduce the data sets and related small target detection algorithms proposed in recent years, and classify, analyze, and compare the corresponding training strategies.

Keywords: Convolutional Neural Network; Residual Network; Image Classification.

1. Introduction

Target detection, which has been developed for about two decades, is a very important core direction of computer vision and has made significant progress in recent years. The main task of target detection consists of two main tasks, i.e., target classification and target localization, i.e., identifying the target in the image and labeling the location information of the target in the image, and the detection task is completed. The traditional steps of target detection are region selection, manual feature extraction, and classifier classification. The common traditional methods are VJ detector, HOG features, variable part model (DPM), etc. The usual method of manual feature extraction is difficult to meet the demand of diverse features of the target, which needs to rely on a lot of experience and expertise, and requires manual resetting of features if a change of application scenario, with poor generalization performance. Therefore, this solution has not been a good solution to the problem of target detection feature extraction. With the rise of deep learning, CNNs have been widely used in target detection tasks due to their powerful feature extraction and fitting capabilities. 2012 Alex Krizhevsky proposed the CNN (Convolutional Neural Networks) model at the ILSVRC competition, which has achieved historic success in the field of image classification. This model has made a historic breakthrough in the field and has significant advantages over traditional methods. In 2014, the VGG model was proposed by the VGG (Visual Geometry Group) group at the University of Oxford at the ILSVRC competition. 2014, GoogLeNet won the ILSVRC competition, and the GoogLeNet model consists of multiple groups of Inception modules. In 2015, Kai-Ming He, Xiang-Yu Zhang, Shao-Qing Ren, and Jian Sun of Microsoft Research proposed the ResNet (Residual Network, or ResNet) model. The model won the championship in ImageNet image classification, image object localization, and image object detection competitions. Therefore, the development of target detection technology has been relatively mature, during which a large number of classical target detection algorithms emerged. Currently, target detection algorithms can be mainly divided into two categories: single-stage and two-stage. The single-stage detection algorithms include YOLO series, RetinaNet, SSD algorithm, etc.; while the two-stage detection algorithms include RCNN, SPP-Net, Fast R-CNN, Faster R-CNN, Mask-RCNN algorithm, etc. Among them, YOLO series algorithms and Faster R-CNN algorithms are practical target detection methods in the industry, although some algorithms have been particularly good for normal targets and have made substantial progress in general-purpose target detection.

The ways of defining small targets can be mainly divided into two categories based on a relative scale and an absolute scale. For the definition based on a relative scale, Chen et al [1] defined small targets as those with a median relative area, the ratio of bounding box area to image area, of all target instances in the same class between 0.08% and 0.58%. For the definition of absolute scale: in the MS COCO dataset [2], a small target is defined as a target with a resolution of less than 32 × 32 pixels. According to the results of MS COCO, a public dataset for target detection, there is a significant difference between large and small target detection in terms of detection accuracy, and the detection accuracy of small targets is only half of that of large target detection. The detection of small targets mainly faces the following challenges: firstly, the problem of poor performance of small target detection has not been completely solved due to poor visual features, lack of sufficient appearance information, and fewer useful pixel points corresponding to them, fewer feature points that can be extracted, and more noise. The research progress for small target detection is also relatively slow; secondly, the large-scale benchmark test datasets for small-size target detection are still not comprehensive enough, and the available datasets cannot support model training for small target detection, nor can they be used as an unbiased benchmark for evaluating algorithms, and there is a lack of

large-scale datasets for small target detection; then, because the general target detection network structure is not applicable to detect small targets. Because the pooling and convolution operations make the features of small targets disappear gradually as the network structure deepens, the information on small targets will be particularly small or even disappear by the end of the classification layer. Finally, in the case of small target aggregation, such as the remote sensing image of a crowd taken by UAV, the crowd is densely piled up together at the edge of the image, and it is very difficult to accurately locate and identify each individual, and there is only one point reflected on the deep feature map after multiple downsampling, which leads to the inability of the model to distinguish the targets. Small target detection has become one of the most challenging tasks in computer vision. Small target detection has a very important practical value in a variety of scenarios such as surveillance anti-theft, UAV scene analysis, infrared weak target detection, pedestrian detection, and self-driving traffic sign detection. It can help people to better protect property security, perform scene analysis, detect infrared weak targets, improve traffic safety, etc.

In recent years, many researchers have proposed many excellent small target detection algorithms concerning network structure, training strategy, data processing, etc. Some processing based on the generic target detection algorithms is performed to adapt to specific small target detection applications. For example, Yaeger et al [3] used data augmentation methods, including distortion and deformation, rotation, and scaling, to enhance handwriting recognition, thus significantly improving its recognition accuracy; Wei Wei et al [4] improved YOLOv3 in aerial target detection by reducing some convolution operations and introducing jump layer connections on the basis of YOLOv3 model, which then ensured the real-time improves the detection accuracy of YOLOV3 model under the guarantee of real-time; Wang Dongli et al [5] went for visual small target detection by feature fusion on SSD model to fuse deep feature information with shallow feature information, and then adjust the prior frame according to the small target size so that the model gets better small target detection capability; XING C et al [6] in 2019 by segmenting the original aerial photography image and using GAN network for super-resolution in order to perform pixel enhancement of the original image, which in turn improves the resolution of small targets; Zhao Pengfei et al [7] the algorithm uses multi-scale null convolution cascade to expand the perceptual range of the feature map and uses 1×1 convolution for image feature information fusion on this basis. The feature maps of corresponding sizes (38 × 38, 10 × 10, and 19 × 19) are then stitched together and channel weighting is implemented using ECAM modules. This design allows for better detail extraction and feature fusion of images and improves the detection accuracy of the algorithm. Woo et al [8] proposed the StairNet algorithm in 2018, which fuses feature information through deconvolution with a shallow layer while passing the fused features to the next deconvolution layer in a top-down manner and then enhances the target semantic information; Li et al [9] proposed an attention mechanism, which they named KNCA-Fusion method tested on public datasets achieved significant results.

2. Small Target Dataset

This subsection describes the dataset used for small target detection. Since deep learning-based target detection algorithms are all data-driven, however, the number of large targets in the initial target detection dataset is much higher than that of small targets, which is an important factor limiting the development of small target detection techniques. Due to the relatively small portion of small targets in the image field of view and the inconspicuous or even missing edge features, deep learning-based small target detection algorithms are ineffective on common target detection datasets because of the limited resolution and amount of information. Therefore, training and detection tasks for small target feature databases are needed. To promote the development of small target detection techniques, many datasets targeting small target detection have been proposed and published in recent years.

2.1. MS COCO Dataset

The COCO [10] dataset is called Microsoft Common Objects in Context (MS COCO), and the COCO dataset is a dataset covering large-scale object detection, segmentation, key-point detection, key-point detection, and captioning tasks. The dataset contains a huge number of 328,000 images. In terms of target detection, a large number of small targets are included. A total of 91 classes of targets are included, with 328,000 images and 2.5 million annotated frames. an example diagram of the COCO dataset is shown in Figure 1. Link to the COCO dataset: https://cocodataset.org/.

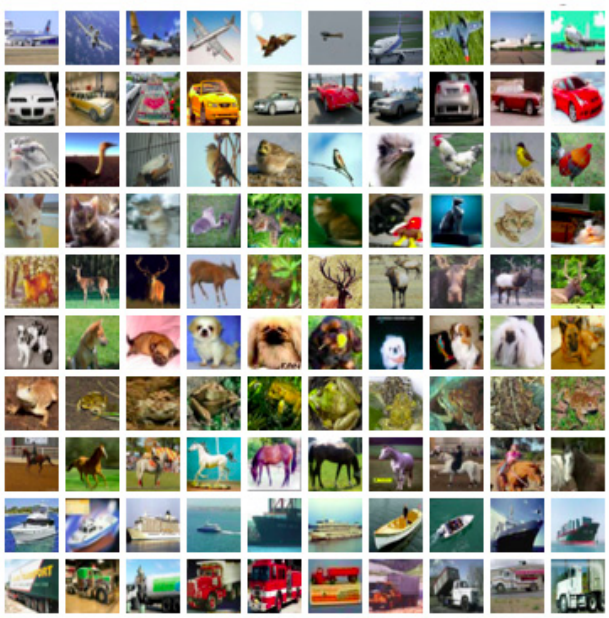

Figure 1. Example graph of the COCO dataset

2.2. EuroCity Persons Dataset

Figure 2. Example graph of EuroCity Persons dataset

The EuroCity Persons [11] dataset focuses on urban traffic scenarios and contains a large variety of accurate and detailed targets. This dataset is almost an order of magnitude larger than the previous dataset used for benchmarking. The dataset covers a large variety of categories and is rich in detail, taking the annotation of people in urban traffic to a new level. EuroCity Persons An example graph of the dataset is shown in Figure 2.

2.3. DOTA Dataset

The DOTA dataset [12] is a large dataset for target detection in aerial images and contains objects of various scales, orientations, and shapes. The annotated DOTA dataset contains 188282 samples for all images. It includes the dataset of remote sensing image target detection area, including 15 categories with 2806 images. the DOTA dataset has been applied to CVPR21 small target detection. the example diagram of the DOTA dataset is shown in Figure 3.

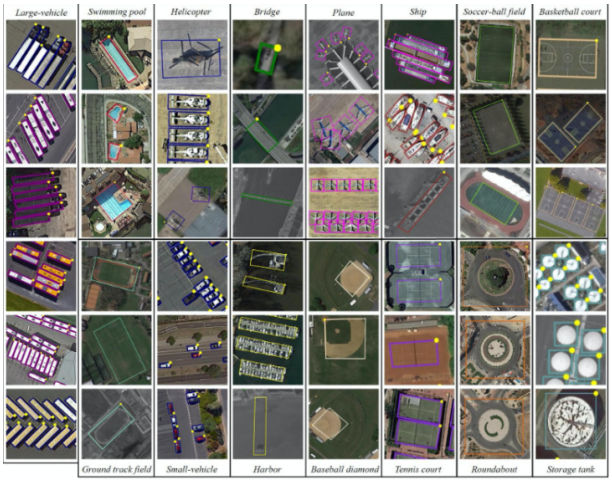

Figure 3. Schematic diagram of the DOTA data set.

2.4. TinyPerson Dataset

The TinyPerson dataset [13] is a benchmark for tiny object detection at long distances and large-scale backgrounds. The images in TinyPerson are collected from the web. First, high-resolution videos are collected from different websites. Second, the images in the videos are sampled every 50 frames. Then images with certain duplicates (homogeneity) are removed and 72,651 objects with bounding boxes are used to manually annotate the resulting images. The training set contains 794 images and the test set contains 816 images. an example graph of the TinyPerson dataset is shown in Figure 4.

Figure 4. Schematic diagram of the TinyPerson dataset.

2.5. Deepscores Dataset

The goal of the DeepScores [14] dataset is to advance the state of the art in small object recognition and to place the object recognition problem in the context of scene understanding. DeepScores contains high-quality sheet music images divided into 300,000 written music sheets containing symbols of different shapes and sizes. To improve the technology of small object recognition and to place the object recognition problem in the context of scene understanding is studied. Datasets that can be used to segment, detect and classify tiny objects. Has been used for road vehicle anomaly detection and for detecting anomalies in video streams. an example graph of the DeepScores dataset is shown in Figure 5.

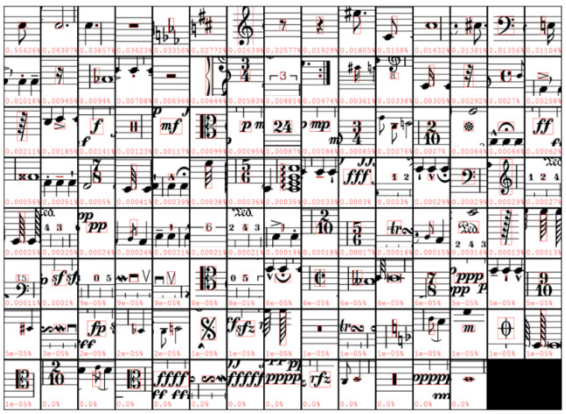

Figure 5. Schematic diagram of the DeepScores dataset

2.6. Other Small Target Datasets

Other datasets include the WIDER FACE [15] dataset, CityPersons [16] dataset, AI-TOD [17] dataset, ISAID [18] dataset, WiderPerson [19] dataset, and Penn-Fudan pedestrian detection and segmentation dataset.

3. Optimization Methods for Small Target Detection

Commonly used optimization algorithms to improve the accuracy of small target detection include multi-scale feature fusion and feature enhancement techniques, a combination between networks, data enhancement, hyperparameter tuning, optimization of backbone networks, the introduction of attention mechanism, optimization of cross-comparison function, optimization of the loss function, an increase of the perceptual field, use of Anchor-Free mechanism, contextual learning, deep learning combined with traditional methods for small target detection, etc. This subsection will introduce several common typical optimization methods related to improving the accuracy of small targets.

3.1. Data Enhancement

Data enhancement mainly includes distortion, rotation and scaling, elastic distortion [20], random cropping [21], panning [22], horizontal flipping, adjusting image exposure and saturation, CutOut [23], Mix Up [24], CutMix [25], Mosaic [26], and other methods. Data augmentation is an effective strategy to solve the problems of small targets with little information and insufficient appearance features and textures to some extent. With data augmentation, the generalization ability of the network can be improved and good results can be achieved in the final detection performance.

Kisantal et al [27] generated images by oversampling these

images and zooming in on each small target, copying and pasting small targets multiple times to increase the number of training samples of small targets, which in turn improves the small target detection performance. Evaluating different pasting enhancement strategies achieved a relative improvement of 9.7% in object detection of small objects on instance segmentation. The Cotout data enhancement proposed by DEVRIES T et al. is also similar to RandomErasing data enhancement by filling the region, thus masking the image information in the filled region, which is beneficial to improve the generalization ability of the model. Unlike Random Erasing, Cutout uses a fixed-size square region with all-0 pixel value padding and allows the square region to be outside the image. ZHANG H [24] proposed Mixup data enhancement, a data enhancement method published in ICLR in 2018, with the core idea of randomly selecting two images from each batch and blending them in a certain ratio to generate new images. The entire training process is trained using only the blended new images, and the original images are not involved in the training process. Cutmix data augmentation proposed by YUN S [28] et al. is to cut out a part of the region but not to fill 0 pixels but to randomly fill the region pixel values of other data in the training set, and the classification results are assigned in a certain proportion. Cutmix simply selects two images from the dataset, and then A part of one image is cropped and superimposed on top of the other image as a new input image into the network for training. Bochkovskiy A [29] et al. proposed a data enhancement method called Mosaic in YOLOV4. The main idea of this method is to crop four images randomly and then stitch them on top of one image as training data. The advantage of Mosaic data enhancement is that it not only enhances the diversity of training data but also enriches the background of the images. The Mosaic data enhancement method is an improvement on the CutMix data enhancement approach and its implementation is also inspired by the CutMix data enhancement approach. MÜLLER S G [30] et al. used all data enhancement approaches to enhance a single image and then sampled uniformly from it. The method is valid for a wide range of datasets and models and is robust, and has been tested on datasets with good result enhancement.

3.2. Introducing the Attention Mechanism

The core idea of Attention Mechanism in computer vision is to find the correlation between data based on the original data and highlight the important features in it. The attention mechanism includes different forms such as channel attention, pixel attention, and multi-order attention. For small target detection tasks, for convolutional neural networks to learn more feature information, the network structure needs to be deepened or widened. However, this would make the neural network model very complex. Meanwhile, the feature information of small targets is weakly expressed because the small targets themselves have less pixel information. Therefore, it is very important to enhance the small target feature information. The introduction of attention mechanisms that can enhance the feature expression ability of small targets, such as channel attention mechanisms and spatial attention mechanisms, often enables neural networks to enhance the feature expression ability of small target information in this way, which in turn improves the detection accuracy of small targets.

The SENet proposed by Jie H [31] considers the relationship between feature channels and incorporates an attention mechanism on the feature channels.SENet is a method to automatically obtain the importance of each feature channel employing learning. For the input feature layer, SENet pays attention to the weight of each channel, and its focus is on obtaining the weight of each channel in the input feature layer. By using SENet to obtain the importance level of each channel, it improves feature representation and suppresses features that are not important for the task at hand, allowing the network to focus on the channels it needs to focus on the most. Li et al [32] proposed a model called YOLO-CAN for small and occluded targets. This model introduces an attention mechanism in its residual structure and improves the feature representation of small target objects by up-sampling and fusing feature maps at different scales. The detector model is inspired by the high detection accuracy and speed of YOLOv3 and is improved by adding an attention mechanism, CIoU (complete intersection on union) loss function, soft NMS (non-maximum suppression), and depth direction separable convolution. PAN H et al [33] proposed a new first-level target detection network, called adaptive dense feature pyramid network (ADFPNet), for detecting targets at different scales. The network was developed on a single-trigger multi-box detector (SSD) framework with a newly proposed ADFP module, which consists of two parts: a dense multiscale and sensory field block (DMSRB) and an adaptive feature calibration block (AFCB). Specifically, the DMSRB block extracts rich semantic information in a dense manner by atrous convolution at different code rates to extract dense features at multiple scales and receptive fields; the AFCB block calibrates dense features to retain features that contribute more and suppress features that contribute less. Tests on many datasets have significantly improved the accuracy and met the requirements for real-time detection. ZHU G [34] proposed a target detection method combining multilevel feature fusion and area channel attention (ODMC). The method first fuses the positive and negative phase information of multi-level features based on CReLU and then uses regional channel attention to further extract target features. The semantics of low-level features and the location information of high-level features are enhanced. Secondly, for the channels after feature fusion, the region information of the feature map is used to optimize the weight assignment, which helps to accurately focus on important channels and suppress irrelevant channels. Finally, the targets are classified and localized based on the enhanced features. The final experiment verifies that ODMC achieves significant improvements and high efficiency on comparable state-of-the-art detection models.

3.3. Feature Fusion

The fusion of features at different scales is one of the important means to improve detection performance. Low-level features have higher resolution and contain more location and detail information, but they are less semantic and noisier due to less convolution being undergone. High-level features have stronger semantic information, but lower resolution and poorer perception of details. How to fuse these two features efficiently to fully utilize their advantages and avoid their disadvantages is the key to improving the segmentation model. There are many studies to improve the performance of detection and segmentation by fusing features from multiple levels. According to the order of fusion and prediction, these methods can be classified as early fusion and late fusion. In complex environments, small targets are often

susceptible to interference from background information. The semantic information extracted by the feature extraction network is relatively limited. In the feature extraction process of target detection, the shallow feature map contains resolution information. Although higher-level features can be used to improve the regression accuracy of the bounding box, these features have little semantic information and are easily disturbed by noise points. The deep-level network contains strong semantic information but has low resolution and the ability to express details. The introduction of feature fusion is effective and can improve the detection of small targets.

SHI W et al [35] proposed an accurate and effective target detection method called feature-enhanced fusion for single-shot target detection (FFESSD), which enhances and utilizes shallow and deep features in the feature pyramid structure of the SSD algorithm. A feature fusion module and two feature enhancement modules are introduced and integrated into the traditional SSD structure. Tests using the proposed network on the dataset show the state-of-the-art mAP, which outperforms the conventional SSD, deconvolution single pass detector (DSSD), feature fusion SSD (FSSD), and other advanced detectors. In extended experiments, FFESSD outperforms conventional SSD for fuzzy target detection. WOO S [36] et al. started from the SSD framework, where the lower layers responsible for small objects lack strong semantics (e.g., contextual information) due to the pyramidal design. This problem was addressed by introducing a feature combination module that unfolds strong semantics in a top-down manner. The final model StairNet detector effectively unifies the multi-scale representation and semantic distribution. Experiments on many datasets show that StairNet significantly improves the weaknesses of SSD and outperforms other state-of-the-art first-class detectors.

3.4. Contextual Information

Contextual information refers to the fact that in an image, a pixel or a target does not exist alone, but has a certain dependency relationship with the surrounding pixels and targets. Reasonably extracting and using the relationship between targets and targets has a great improvement on the detection accuracy of small targets.

TANG X et al [37] proposed a new context-assisted single-shot face detector, called a pyramidal box, for dealing with difficult face detection problems. Due to the importance of context, the utilization of contextual information is increased at three levels as follows. First, we propose a novel contextual anchor that employs a zero-point-five supervised approach to track the learning of the environment's high-level context, called the pyramid anchor. Second, a low-level feature pyramid network is proposed to combine sufficient high-level contextual semantic features with low-level facial features, which also allows pyramid boxes to predict faces at all scales in a single shot. Third, a context-sensitive structure is introduced to increase the capacity of the prediction network to improve the accuracy of the final output. HU H et al [38] proposed an object relationship model. The model takes advantage of the appearance properties of objects, and the interactions between geometries to process each object simultaneously, thus allowing the modeling of the relationships between them. It is lightweight and in situ. The model requires no additional supervision and is easily embedded in existing networks. It proves to be effective in improving object recognition and repetition removal steps in modern object detection pipelines. The model validates the effectiveness of modeling object relationships in CNN-based detection. The model yields the first fully end-to-end object detector.

4. Conclusion

Many methods and techniques have been proposed for improving the detection accuracy of small targets, and this paper has done a review of a few simple techniques. This paper first starts with the development of target detection briefly describes the obstacles to the development of target detection, and then transitions to small target detection. Secondly, the common datasets for small target detection are introduced. Finally, optimization methods for small target detection are described, and several classical optimization methods for small target detection are selected and elaborated.

The current network model still has a complex network structure model and too many parameters of the network model, which are difficult to deploy. The detection accuracy for small targets is low, so it still needs a long time of development and research to improve the small target detection technology. For data enhancement methods, first of all, when data enhancement also has to consider whether the problem of introduced noise points will affect the learning of small target features by neural networks and other issues. Second is also to consider the network structure of different networks combined to improve the detection accuracy of small targets, super-resolution reconstruction is one of the most direct and interpretable methods to improve the performance of small target detection. A feasible future research idea is to deeply combine the advanced technology of super-resolution reconstruction with target detection technology. Lastly, it is important to improve the small target data set as much as possible, because data enhancement also has certain limitations, and only enough data is an important cornerstone to improve the small target detection accuracy.

References

[1] CHEN C, LIU M Y, TUZEL O, et al. R-CNN for small object detection[C]//Proceeding of Asian Conference on Computer Vision. Cham: Springer, 2016: 214-230.

[2] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[3] YAEGER L, LYON R, WEBB B.Effective training of a neural network character classifier for word recognition[J].Advances in Neural Information Processing Systems,1996,9: 807-816.

[4] Wei Wei, Pu Wei, Liu Yi. Improvement of YOLOv3 in aerial target detection [J]. Computer Engineering and Applications, 2020, 56(7): 17-23.

[5] Wang Dongli, Liao Chunjiang, Mou Jinzhen, et al. Feature fusion-based small target detection for SSD vision [J]. Computer Engineering and Applications, 2020, 56(16): 31-36.

[6] XING C, LIANG X, BAO Z.A small object detection solution by using super-resolution recovery[C]//2019 IEEE 7th International Conference on Computer Science and Network Technology(ICCSNT),2019：313-316.

[7] Zhao P.F., Xie L.B., Peng L. A deep small target detection algorithm incorporating attention mechanism [J/OL]. Computer Science and Exploration [2021-10-18].

[8] WOO S, HWANG S, KWEON I S. StairNet: top-down semantic aggregation for accurate one-shot detection [C]// Proceedings of the 2018 IEEE Winter Conference on

Applications of Computer Vision, Lake Tahoe, Mar 12-15, 2018. Piscataway: IEEE, 2018:1093-1102.

[9] Li W.T., Peng L. Small target detection algorithm for multi-scale channel attention fusion networks[J]. Computer Science and Exploration, 2021, 15(12): 2390-2400.

[10] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context［C］// European Conference on Computer Vision. Zurich: ECVA,2014: 740-755.

[11] Pang Y, Cao J, Li Y, et al. TJU-DHD: A Diverse High-Resolution Dataset for Object Detection[J]. IEEE Transactions on Image Processing, 2021, 30:207-219.

[12] XIA G S,BAI X,DING J,et al. DOTA: a large-scale dataset for object detection in aerial images［C］ //IEEE Conference on Computer Vision and Pattern Recognition.Piscataway: IEEE, 2018: 3974-3983.

[13] YU X,GHONG Y,JIANG N,et al. Scale match for tiny person detection ［C］//Winter Conference on Applications of Computer Vision.Piscataway: IEEE,2020: 1246- 1254.

[14] Tuggener L, Elezi I, Schmidhuber J. et al. DeepScores -- A Dataset for Segmentation, Detection and Classification of Tiny Objects: IEEE Computer Society, 10.1109/ ICPR. 2018. 8545 307 [P]. 2018.

[15] YANG S, LUO P, LOY C C, et al. WIDER FACE: a face detection benchmark[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27- 30, 2016. Washington: IEEE Computer Society, 2016: 5525-5533.

[16] ZHANG S, BENENSON R, SCHIELE B. CityPersons: a diverse dataset for pedestrian detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017:3213-3221.

[17] WANG J W, YANG W, GUO H W, et al. Tiny object detection in aerial images[C]//Proceedings of the 2020 25th International Conference on Pattern Recognition, Milan, Jan 10-15, 2021. Piscataway: IEEE, 2021: 3791-3798.

[18] ZAMIR S W, ARORA A, GUPTA A, et al. iSAID: a large-scale dataset for instance segmentation in aerial images [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, Jun 16- 20, 2019. Piscataway: IEEE, 2019: 28-37.

[19] ZHANG S F, XIE Y L, WAN J, et al. WiderPerson: a diverse dataset for dense pedestrian detection in the wild [J]. IEEE Transactions on Multimedia, 2020, 22(2): 380-393.

[20] SIMARD P Y, STEINKRAUS D, PLATT J C. Best practices for convolutional neural networks applied to visual document analysis [C]//Proceedings of ICDAR. [S.l.]: IEEE, 2003, 3 (2003).

[21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems,2012,25: 1097-1105.

[22] WAN L, ZEILER M, ZHANG S, et al. Regularization of neural networks using dropconnect[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2013: 1058-1066.

[23] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[EB/OL]. (2017-08-15)[2017-11-29].https://arxiv.org/abs/1708.04552.

[24] ZHANG H, CISSE M, DAUPHIN Y N,et al.Mixup: Beyond empirical risk minimization [EB/OL]. (2017-10-25) [2018- 04-27].

[25] Yun S. Han D. Oh S J. et al. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features[J]. 2019.

[26] Bochkovskiy A, Wang C Y, Liao H . YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.

[27] KISANTAL M, WOJNA Z, MURAWSKI J, et al. Augmentation for small object detection[EB/OL]. (2019-02-19) [2019-02-19].

[28] YUN S, HAN D, OH S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 6023-6032.

[29] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2020-04-23].https://arxiv.org/abs/2004.10934.

[30] MÜLLER S G,HUTTER F.TrivialAugment：tuning- free yet state- of- the- art data augmentation[J].arXiv：2103. 10158, 2021.

[31] Jie H, Li S, Gang S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[32] LI Y, LI S, DU H, et al. YOLO- ACN: focusing on small target and occluded object detection[J]. IEEE Access, 2020, 8: 227288-227303.

[33] PAN H, CHEN G, JIANG J. Adaptively dense feature pyramid network for object detection[J]. IEEE Access, 2019, 7: 81132-81144.

[34] ZHU G, WEI Z, LIN F. An object detection method combining multi-level feature fusion and region channel attention[J]. IEEE Access, 2021, 9: 25101-25109.

[35] SHI W, BAO S, TAN D. FFESSD: an accurate and efficient single-shot detector for target detection[J]. Applied Sciences, 2019, 9(20): 4276.

[36] WOO S, HWANG S, KWEON I S. StairNet: top-down segmantic aggregation for accurate one-shot detection[C]// Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, Mar 12-15, 2018. Piscataway: IEEE, 2018:1093-1102.

[37] TANG X,DU D K,HE Z,et al. PyramidBox: a context assisted single shot face detector [C]// European Conference on Computer Vision. Zurich: ECVA,2018: 812-828.

[38] HU H,GU J,ZHANG Z,et al.Relation networks for object detection [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2018: 3588-3597.