International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

A Self-Learning Object Detection Method Based on Highly Reliable Sample Mining

Di Li[1], Dongshan Li[2], W. Ni[3]
[1,2] College of Information Engineering, Henan University of Science and Technology, Luoyang, China.
[3] Fudan Institute on Networking System of AI, Shanghai, China.

ABSTRACT: Visual object detection is an artificial intelligence technique that locates specific objects from images, which is of great significance for practical applications. However, training general object detection models require many manually annotated images, bringing more labour and time cost. In order to improve the adaptability of the object detection model to the data environment changes, this paper proposes a self-learning object detection system based on high-reliability sample mining. We first train a SampleNet that can better mine reliable training samples from unlabeled data. We then use the combination of SampleNet and the basic object detection model to build a complementary residual training framework, continuously improving the sample mining ability and object detection tasks during the training process. The experimental results show that SampleNet can stably provide reliable pseudo samples for model training, and the complementary residual training framework improves the performance of basic object detection tasks.

KEYWORDS: Object detection, Complementary residual learning, Sample mining.

I. INTRODUCTION

Object detection [1] is a computer vision technique that extracts semantic object instances from digital images and videos. It is widely used in object segmentation, image retrieval, video surveillance, spatial understanding and automation driving, target tracking, face detection and other application fields. Relying on the research of various deep learning models based on convolutional neural networks [2], the performance and accuracy of object detection systems have made significant progress in recent years. In the detection tasks of datasets such as Pascal VOC [3], object detection models with different architectures constantly refresh the optimal records and have achieved good application results in some engineering fields. However, the current visual object detection system mainly relies on fixed manual annotation datasets for model training, which requires a lot of labour and time to collect training data. At the same time, the training and application of current object detection models are mainly based on cloud computing platforms. For complex and diverse application scenarios, the adaptability of the unified platform model will be challenged [4].

With the rapid development of communication and computing fusion technology, especially the enhancement of computing and communication capabilities of edge terminal devices, the cloud-centric unified object detection model will not be able to meet the increasing data processing needs. Massive new unlabeled data often conflict with the feature distribution learned by the cloud unified model, causing the generalization of the fixed target detection model to decrease gradually over time. In order to improve the adaptability of the object detection system to the environment, the key technical challenge is to strengthen the self-learning ability of the target detection model for perceptual information to realize the rapid iteration and deployment of the model.

In this paper, we propose a self-learning object detection system based on high-reliability sample mining, which effectively solves the problem of model adaptation to data changes. Specifically, we modify the confidence loss calculation method of the general object detection model and propose the SampleNet method based on yolov3. The training model extracts more reliable target samples from unlabeled data for self-training by suppressing the model's low confidence predicted bounding boxes. Then, we jointly train SampleNet and the basic object detection model and propose a complementary residual training framework, which effectively utilizes the highly reliable pseudo-labels of SampleNet and continuously improves the performance of the overall object detection system. Experimental results on multiple datasets show that our proposed object detection method can better adapt to changes in training data and achieve better model performance through self-learning.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

II. RELATED WORK

For the efficient and autonomous optimization of visual object detection models while reducing human intervention and improving data utilization efficiency, the main goal of current related research is to achieve direct or indirect pseudo-label acquisition of new data. Pseudo-labels will be used for direct or assisted training of the model. According to the task's complexity and the degree of automation, the main methods in the current research field include active learning, weakly supervised and self-supervised learning [5].

A. Object Detection Based on Active Learning

The main goal of such methods is to design a reliable example mining algorithm [6] to obtain target instance labels from unlabeled data that can be directly used for training. These methods introduce an active learning framework in the sample mining process and manually annotate the indistinguishable predictions [7]. The method based on active learning and sample mining proposes better evaluation and selection criteria for high-confidence samples, which reduces the object detection model's dependence on the manual annotation to a certain extent. However, there is still a considerable gap from automatic model training.

B. Object Detection Based on Weakly Supervised Learning

The weakly supervised object detection model aims to achieve better object detection performance in the case of sparse or inaccurate sample labels. [8] propose two methods for sample acquisition, using the partial perceptual sampling method to traverse the region candidates to obtain categories that are not related to the true labels, and discard the loss of these categories during the training process of the region candidates. [9] propose a method based on detection difference for weakly supervised semantic segmentation; [10] propose a progressive approach to train the weakly supervised semantic segmentation model to achieve layer-by-layer progressive optimization training. [11] propose a object detection framework from weak supervision to full supervision, using an iterative learning algorithm to continuously correct pseudo-labels until the pseudo-label correction results were optimal. Other researchers achieve weakly supervised learning for object detection through saliency map-based image representation. Xie et al. propose a two-stage cascaded CNN structure to learn the features of the target adhesion region, which improved the multi-object detection performance of the model. The object detection model based on weakly supervised learning has achieved better target localization or detection performance in sparse samples. The proposed idea, such as sample acquisition, difference judgment, and alternate training, have high reference values. However, because the accuracy of sample acquisition is limited by specific tasks, the task generalization is insufficient.

C. Object Detection Based on Self-supervised Learning

The detection method based on self-supervised learning mainly studies the algorithm of directly obtaining the target pseudo-label to realize the autonomous training and iteration of the whole process, and it pays more attention to the optimization of the automatic object detection model. [12] uses the characteristics of solid continuity of background data to mine samples by predicting the background. Pathak et al. automatically predicts foreground targets in unlabeled videos based on the principle of strong target motion, which better achieves target prediction in single-frame static pictures. [13] adopt a domain adaptation method for automatic object detection, and used generative adversarial networks to perform style transfer from source data to target data. [14] propose a method of interactively training classifiers and detectors, using the context information of the previous model as an aid to training the current model. [15] apply self-supervised learning to label reuse tasks, combining multi-task learning with self-supervised learning, including the main task and multiple auxiliary tasks. The object detection model based on self-supervision adopts a variety of direct pseudo-label acquisition methods according to the characteristics of the task and optimizes the model in an intuitive process. These methods focus on correcting the accuracy of the object detection model, which promotes the use of unlabeled data for training. The ideas of label reuse and auxiliary tasks in the above methods provide a good reference for this research.

III. PROPOSED METHOD

A. High-confidence sample mining model

Generally, the single-stage object detection model has a high architecture integration degree, and the judgment and calculation of the model output are more complicated. In order to train a novel model structure suitable for sample mining tasks using single-stage object detection methods, adjustments to the underlying loss function need to be considered. Among the multiple loss functions of Yolov3 [16], the confidence loss for judging whether the predicted bounding box contains objects is determined according

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

to the intersection over union (IOU) of the ground truth and the predicted bounding box. In the model training process, when calculating the confidence loss, the model will get the predicted bounding box which has the largest IOU with the ground truth box and set its confidence to be 1. The predicted boxes whose IOU with the ground truth box is less than the confidence threshold will be set to 0. In addition, the rest of the predicted boxes will be ignored. The function of this setting is to ensure the detection rate of the model.

We aim to mine more valid examples from the image, and hope that the target bounding box detected by the model has higher confidence and reliability. Therefore, during the model training phase, we change the calculation of the confidence loss. Specifically, we only use the bounding box predicted by the model have the largest IOU with the ground truth box as the positive sample for confidence loss calculation, and use all the remaining predicted boxes as negative samples for training. This method significantly enhances the reliability of the model for discriminating samples. Still, since the number of negative samples is much larger than that of positive samples, there may be a class imbalance problem during training. To alleviate this problem, we replace the general binary cross-entropy loss with Focal loss. The improved confidence sample mining model is shown in Figure 1.

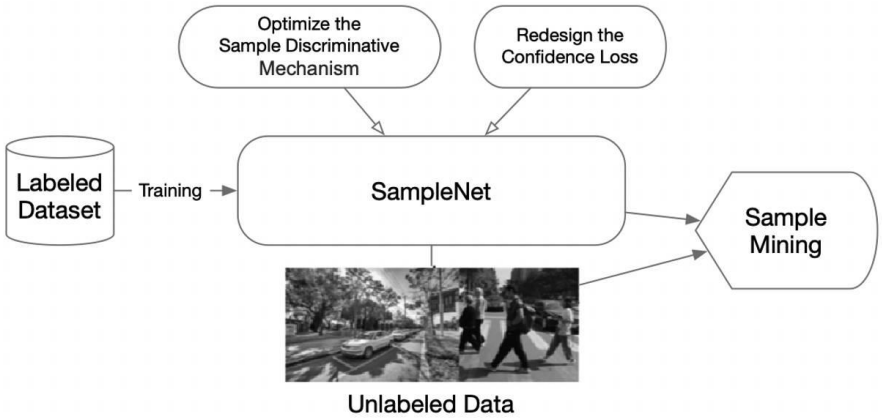

Figure 1. High-confidence sample mining model.

B. Complementary residual model

We build an overall training framework that combines SampleNet and a base object detection model. The proposed framework adopts a dual-model synchronous joint training approach. Using the characteristics of high output confidence and low recall rate of the sample mining model, combined with the characteristics of the basic target detection model's low output accuracy and high recall rate, a complementary residual model for learning the difference between the two outputs is established. The main purpose of this framework is to optimize the basic object detection model, and the complementary training and collaborative optimization of the two models are realized.

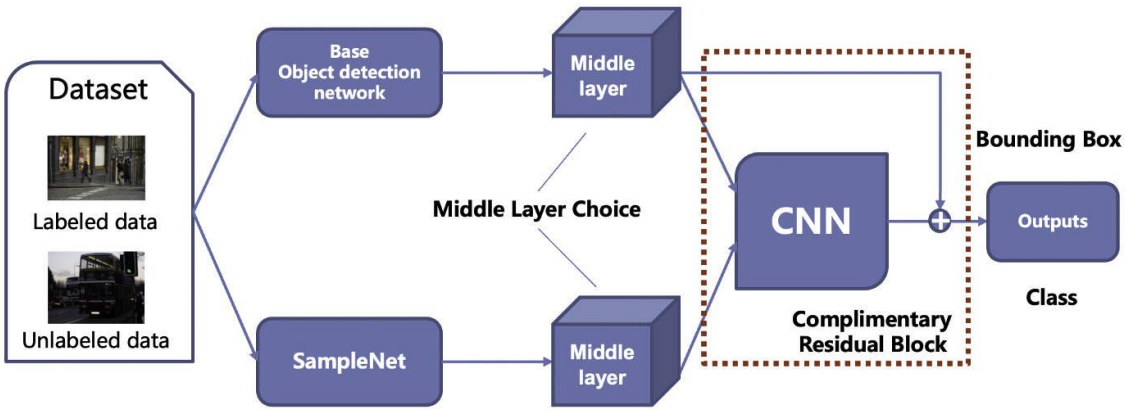

Figure 2. Illustration of the complementary residual model

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

The overall structure of the complementary residual model is shown in Figure 2. SampleNet mainly deals with unlabeled data, while the basic object detection model primarily uses labelled data for supervised training. In the joint training framework, the feature layer with higher expressive ability will be used to calculate the complementary residual loss after the middle layer selection. The residual module ensures the information exchange between the two models, so that the SampleNet and the basic detection model can continue to obtain useful complementary information for training. The highly reliable predicted bounding box obtained by SampleNet will also be used to train the basic object detection model, thereby continuously improving the model accuracy. The proposed framework improves the recall rate of SampleNet and the accuracy rate of the basic target detection network synchronously, so as to better realize the self-learning and adaptive ability of the model to unlabeled data.

IV. EXPERIMENTS

A. Experiment process of the high-confidence sample mining model

For the high-confidence sample mining model, we improve the output decision mechanism based on the single-stage high-performance YOLOv3 object detection framework. In view of the significant difference in the number of positive and negative samples caused by the modification of the target confidence judgment mechanism, we introduce a new confidence loss judgment function, conduct model optimization training on multiple data sets, and verify the test results on unlabeled data to evaluate the effectiveness of SampleNet.

The specific experimental process is as follows: 1) We use the Pascal VOC 2007 and 2012 data sets for model training and experimental validation; among them, the training set is all the data of the VOC2007 data set, and the validation set is the training set of VOC2012, and the repeated validation set is VOC 2012 validation set. 2) Design a sample mining model, modify the model's target confidence determination mechanism and corresponding loss function, and adjust hyperparameters through experiments. 3) Put the training set into the new design model for iterative training, obtain the trained SampleNet, and use the training set to train the basic detection model for subsequent experimental verification. 4) In the experimental validation stage, firstly, the sample mining model is used to output samples on the verification set, and the difference between the samples and the ground truth labels is judged. The goal is to make the obtained samples approach the accuracy of the ground truth labels. Then, the pseudo-labels obtained by SampleNet are used as training samples to optimize the object detection model, so that the mean average precision (mAP) of the object detection model can approach or exceed the level of training with ground truth labels.

B. Experiment process of the complementary residual model

For the complementary residual model, we create a basic object detection model and combine it with SampleNet to design a synchronous joint training mechanism. By judging the middle layer with higher feature representation ability in the two network models, a complementary residual module is established to evaluate the output difference between the two models. With the main goal of optimizing the basic object detection model, the synchronization optimization of the two models is realized.

The specific experimental process is as follows: 1) Remove half of the actual annotations from the training set of the VOC 2012 dataset as the training set of this experiment. 2) Build a basic target detection model and combine it with SampleNet. 3) Select the middle layer with better feature representation in the two models through multiple training and experiments, and design a complementary residual module. Then we measure the difference between the original output of the base object detection model and the joint output after adding supplementary information. 4) Combine the basic model, sample mining model and residual module to form a complementary residual model framework. During the training process, the labelled and unlabeled training data are input alternately, and the joint error is set as the overall loss function to complete the model training. 5) Verify the performance of the basic target detection model and the sample mining model on the validation set to ensure that the performance of the two models can be jointly improved.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

C. *Experimental results*

Figure 3. Example of the prediction results of the basic object detection model

Figure 3 shows the predicted bounding box of the basic target detection model. The blue box in the figure represents the ground truth result. It can be seen that the performance of the basic detection model is weak, and the predicted bounding box is generally quite different from the ground truth.

Figure 4 shows the sample mining results of SampleNet. By suppressing low-confidence prediction boxes, the model can more accurately extract reliable target samples from the input image and use them for training the basic target detection model. Although a small number of objects may not be detected, SampleNet ensures the reliability of detected objects as much as possible so as to avoid the training of the model being disturbed by wrong predictions.

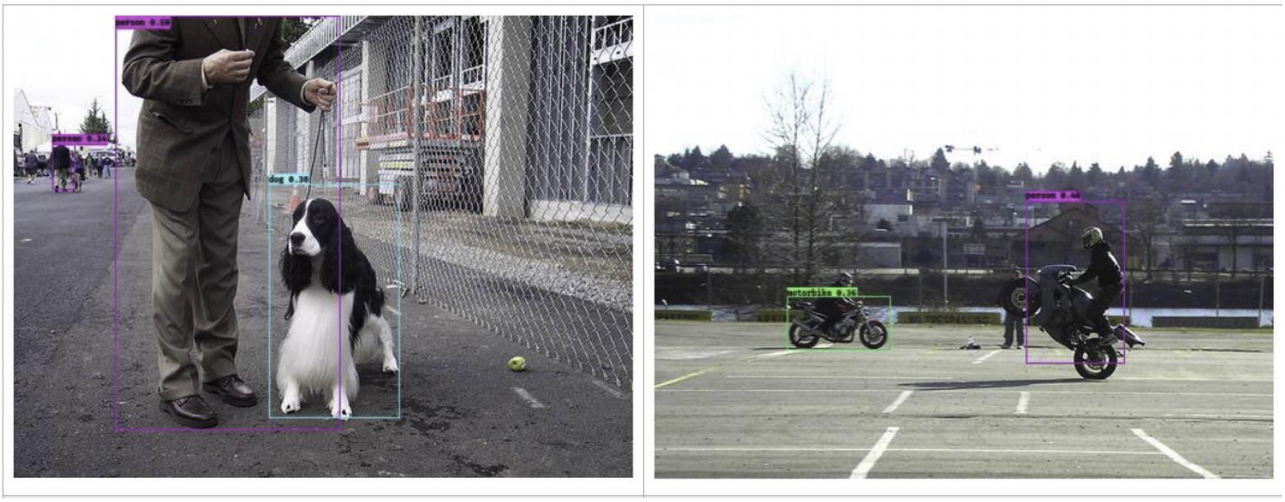

Figure 4. Sample mining results of the trained SampleNet

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

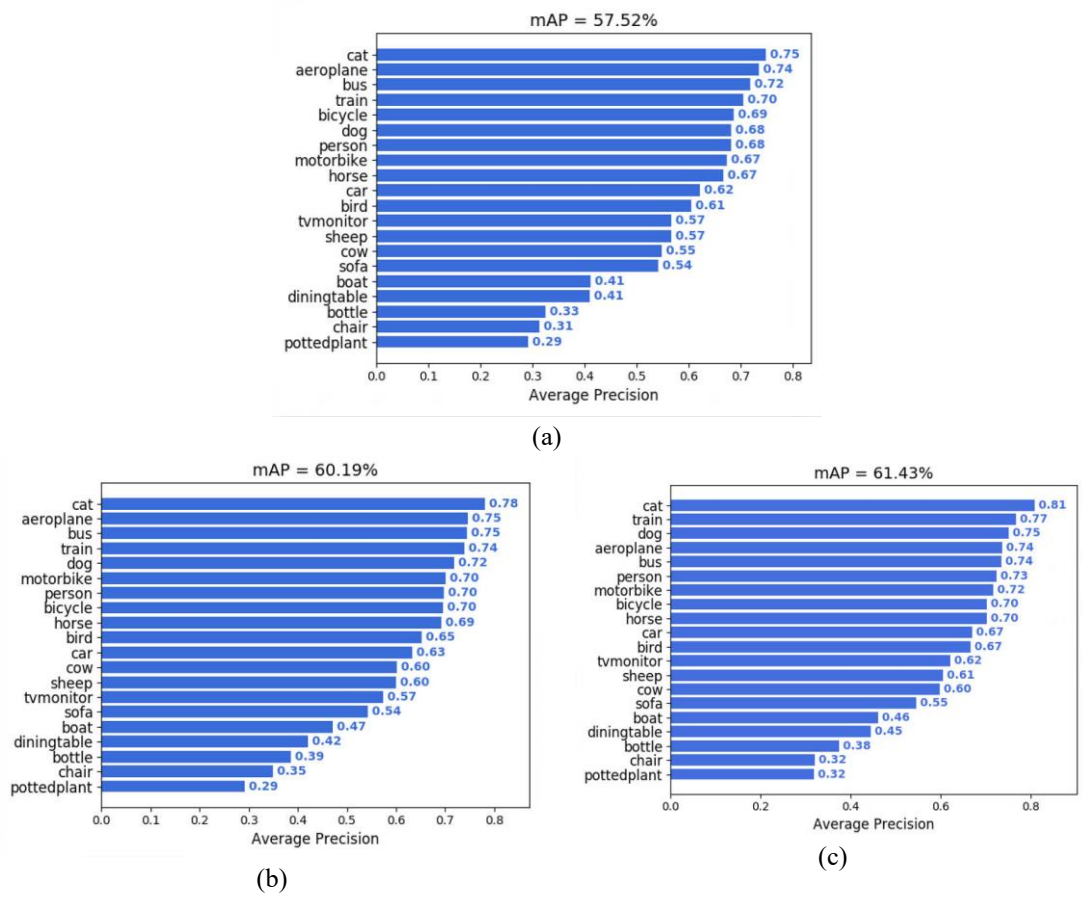

Figure 5. Model performance comparison under different training configurations.
(a) Average precision of the base model.
(b) The average precision of the model after sample mining and retraining based on SampleNet.
(c) Average precision of the complementary residual model.

We compared model performance under several training configurations. As can be seen in Figure 5(a), after the base object detection model is trained on the VOC 2007 dataset, the mAP of the base model on VOC 2012 is 57.52%. Figure 5(b) shows the performance after training the SampleNet to mine valid samples in the VOC 2012 training set and retraining the base model. The mAP of the model on VOC 2012 is improved to 60.19%. Figure 5(c) shows the joint training results of the base model and SampleNet under the complementary residual model framework. The base detection model achieves a better mAP on the validation set. Through the effective combination of SampleNet and the basic object detection model, the proposed complementary residual model framework can utilize a small amount of labelled data for more reliable model self-training and achieve better performance.

V. CONCLUSION

In order to solve the problem that the object detection task is highly dependent on manual annotation and cannot effectively utilize the newly added data, this paper first proposes a highly reliable sample mining model SampleNet. Then, by constructing a complementary residual model combining SampleNet with the basic object detection model, the continuous self-optimization of the two models is realized using only a small amount of labeled data. The experimental results show that SampleNet can obtain effective bounding boxes more reliable from unlabeled data for the training of object detection models. The complementary residual model makes up for the shortcomings of SampleNet and basic object detection models and further optimizes the detection performance. The proposed method provides a technical reference for object detection application in practical scenarios.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 08 August 2022
DOI: 10.47191/ijcsrr/V5-i8-52, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

REFERENCES

1. Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.
2. Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. Econometrica, 89(1), 181-213.
3. Hoiem, D., Divvala, S. K., & Hays, J. H. (2009). Pascal VOC 2008 challenge. World Literature Today, 24.
4. Li, H., Ota, K., & Dong, M. (2018). Learning IoT in edge: Deep learning for the Internet of Things with edge computing. IEEE network, 32(1), 96-101.
5. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
6. Canévet, O., & Fleuret, F. (2015, February). Efficient sample mining for object detection. In Asian Conference on Machine Learning (pp. 48-63). PMLR.
7. Wang, K., Yan, X., Zhang, D., Zhang, L., & Lin, L. Towards Human-Machine Cooperation: Self-supervised Sample Mining for Object Detection.
8. Niitani, Y., Akiba, T., Kerola, T., Ogawa, T., Sano, S., & Suzuki, S. (2018). Sampling Techniques for Large-Scale Object Detection from Sparsely Annotated Objects. arXiv preprint arXiv:1811.10862.
9. Shimoda, W., & Yanai, K. (2019). Self-Supervised Difference Detection for Weakly-Supervised Semantic Segmentation. arXiv preprint arXiv:1911.01370.
10. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M. M., Feng, J., ... & Yan, S. (2015). STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation. arXiv preprint arXiv:1509.03150.
11. Zhang, Y., Bai, Y., Ding, M., Li, Y., & Ghanem, B. (2018). Weakly-supervised object detection via mining pseudo ground truth bounding-boxes.
12. Katircioglu, I., Rhodin, H., Constantin, V., Spörri, J., Salzmann, M., & Fua, P. (2019). Self-supervised Training of Proposal-based Segmentation via Background Prediction.
13. Kim, S., Choi, J., Kim, T., & Kim, C. (2019). Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection. arXiv e-prints, arXiv-1909.
14. Song, Z., Chen, Q., Huang, Z., Hua, Y., & Yan, S. (2011). Contextualizing object detection and classification.
15. Lee, W., Na, J., & Kim, G. (2019, June). Multi-Task Self-Supervised Object Detection via Recycling of Bounding Box Annotations. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4979-4988). IEEE Computer Society.
16. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Cite this Article: Di Li, Dongshan Li, W. Ni (2022). A Self-Learning Object Detection Method Based on Highly Reliable Sample Mining. International Journal of Current Science Research and Review, 5(8), 3220-3226