An Enhanced Light Object Detection for Indiscernible Object in the Special Scene

DOI : 10.36909/jer.ICCSCT.19479

Quanyou Zhang[1,2], Yong Feng[1*], Yong-heng Wang[3], Bao-hua Qiang[4], Lufeng Wang[5], Zebin Zhang[6]

[1]Chongqing University, Chongqing 400044, China
[2]Xuchang University, Xuchang 461000, China
[3]8# of Zhejiang Lab, Yuhang district, Hangzhou 311121, China
[4]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[5]Chongqing Industry Polytechnic College, Chongqing 401120, China
[6]Zhengzhou university, Zhengzhou City, 450001, China

* Corresponding Author: fengyong@cqu.edu.cn

ABSTRACT

We present a transfer learning method named Special Application Transfer (SAT) for special object detection in a real life scenario. Our method improves fine-tuning hyper-parameter and adds unrecognized samples to detect special samples when training object detection neural networks for classification. We implement the model of NanoDet on special supervised datasets and fine-tune the hyper-parameter on a target task. More importantly, we combine a few carefully selected samples in training and simple heuristic fine-tuning to achieve good performance on special object detection in real-life scenarios. Our method (SAT) performs well across surprisingly the small dataset the medium dataset and the large dataset. SAT achieves 95% AP (Average Precision) on the small dataset, 94.8% AP on the medium dataset, and 94.5% AP on the large dataset. The performances of AP run-time and training convergence are perfect, compared with the original method and well-established famous methods on the challenging COCO dataset and our dataset. We hope our work could promote and complete the practical application in more real life scenarios. Our code is available at: https://github.com/zhangquanyou/SAT.

Keywords: Object detection; Hyper-parameter; Transfer learning; Mobile terminal; Algorithm

INTRODUCTION

In computer vision, modern mainstream object detection(Piórkowska-Kurpas et al.,2021) includes the one-stage or the two-stage, predefine marks on a dense feature map grid and detects candidates boxes with the corresponding categories(Qi et al.,2022). In the one-stage method, detectors(Kolesnikov et al.,2019) are predefined marks or reference points(Sun et

al.,2020) on a dense feature map grid to predict box categories of objects. The familiarized one-stage object detection algorithm includes OverFeat, SSD, RetinaNet, YOLOX(Ge et al.,2021), YOLOv1-v7(Junos et al.,2021), and so on. In the two-stage(Zhao et al.,2022) method, a small set of candidate objects are pipelined work on the predicted proposal boxes to classify and identify the target object by the algorithm. There are the familiarized two-stage(Tk et al.,2020) algorithms, such as R-CNN(Wu et al.,2020), SPP-Net, Faster R-CNN(Avola et al.,2021) and R-FCN. However, these models are relatively large and not suitable for transplantation to mobile terminals or embedded devices. In real life scenarios, objects are varied and unpredictable in most application scenarios, on account of different structures, colors and intensity. It is not always accurate, for influencing factors can reduce predicted accuracy. In these cases, we can use fine-tuning(Wang et al.,2021) arguments or innovative approaches to improve the accuracy of object detection, as shown in Figure 1.

Figure 1. Salient object detection.
(The first picture is a demo of side face detection. The second picture is a demo of action recognition. The third picture is a demo of different category detection. The last is a demo of detecting moving vehicles.)

In this paper, we propose a transfer learning method named Special Application Transfer (SAT) for special object detection. We implement the model of NanoDet(Zhou et al.,2021) by adding unrecognized samples to detect special samples and fine-tune hyper-parameters of neural networks for classification. The novelty of our work mainly includes three aspects. First, we improve the method of calculating the loss of IoU, which uses two points coordinate instead of the original algorithm's four points coordinate, as shown in Figure 2. Second, our method combines a few carefully selected unrecognized samples and NanoDet model to train our dataset and COCO dataset. Third, we fine-tuned the hyper-parameter to achieve good performance on special object detection in the real application scenario.

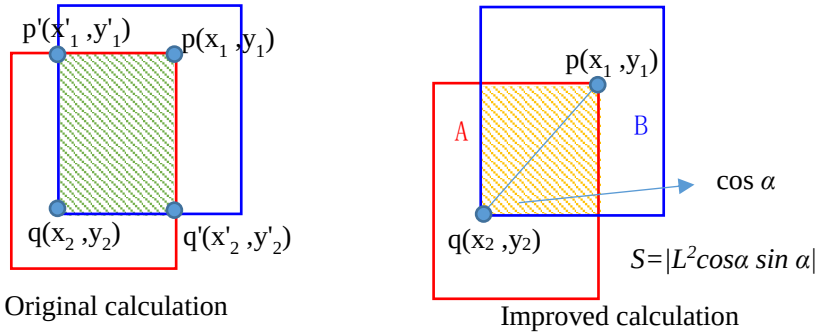

Figure 2. The original and improved calculation of the loss of IoU: A is a predicted box, and

B is the verified ground box.

(Calculating the loss of IoU by calculating area, as shown in the enhanced algorithm 2.)

SAT demonstrates across surprisingly the small dataset the medium dataset and the large dataset. SAT achieves 95% AP (Average Precision)(Tychsen-Smith et al.,2017) on the small dataset, 94.8% AP on the medium dataset, and 94.5% AP on the large dataset. The performances of AP run-time and training convergence are perfect, compared with original method and well-established famous methods on the challenging COCO dataset and our dataset. We hope our work could promote and complete the practical application in more real life scenarios.

Paper Organization: In section 1, we introduce the classification of the familiarized object detection algorithms. In section 2, we show the related work and theory of object detection of SAT. In section 3, we demonstrate the novelty of the paper and the detail of the improved algorithm. In section 4, we show the processing of the experiment and fine-tuning of SAT. In section 5, we compare SAT with NanoDet and discuss the trends of processing time (Forward, Decode and Visualization). In the last section, we show the limitations and future perspectives of our method.

MATERIALS AND METHODS

2.1 Backbone

Most object detection model architecture is composed of three parts: backbone(Chen et al.,2019), task head and connector. We implement SAT model to predict the target object based on NanoDet, which adopted Feature Pyramid Network (FPN)(Roh et al.,2021) of ShuffleNetV2 architecture as the backbone network, as shown in Figure 3. SAT constructs the pyramid with levels to produce multi-scale feature maps from the input image. All pyramid levels have the same in-channels, out-channels, kernel-size, stride, padding, dilation, bias and normalization same as the torch.nn.Conv2d. We align the fine-tuning with NanoDet to show the speed and effectiveness of SAT.

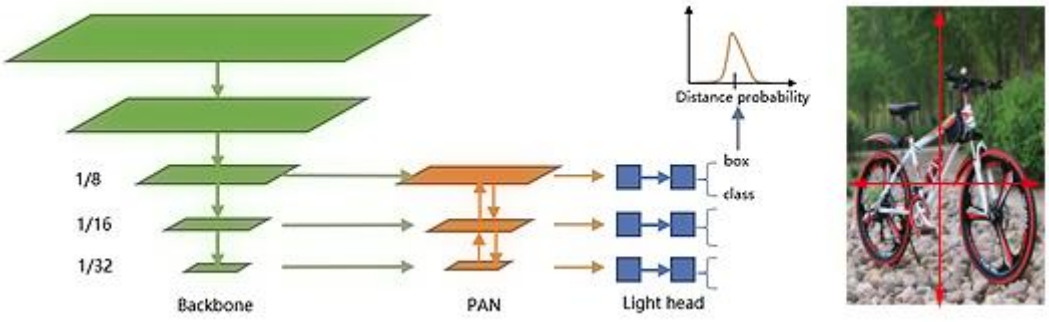

Figure 3 . Architecture of SAT.

(It contains three parts of backbone, pan and light head.)

2.2 Pyramid Attention Network

Pyramid Attention Network (PAN)(Zhao et al.,2021) is an improvement of FPN, which

up-down sampling high-level semantic information. The main function of PAN is to complement FPN and enhance the positioning information of the object. PAN adopts bilinear interpolation(Zilberstein et al.,2021) instead of pooling for up-down sampling, which can effectively reduce the computation. SAT samples backbone layers to output FPN2, FPN3 and FPN4 to construct Pyramid Attention Network (PAN). The model size and activation function of backbones can be modified to form different special application networks. In PAN of SAT, the list of feature map channels extracted from the backbone is [116, 232, 464], and the output is 96 feature map channels.

2.3 Light head

Fully Convolutional One-Stage (FCOS)(Wang et al.,2021) Object Detection series uses detection headers with shared weights to generate prediction boxes for the multi-scale Feature Map. Shared weights are used as the coefficient for each layer to scale the predicted boxes, so as to reduce the number of parameters. Headers utilize the RoIAlign(Xiying et al.,2022) operation to extract features for each box to generate a feature head with its box location and classification. FCOS/ATSS (Adaptive Training Sample Selection)(Huang et al.,2021) is currently used to estimate the quality of detection boxes The GIoU (Generalized Intersection over Union)(Zhang et al.,2022) is also used in some other similar works, which makes up GFL loss scores with the following equation.

$$\mathcal{L}_{GFL} = \frac{1}{N_{POS}} + \sum_z \mathcal{L}_{QFL} + \frac{1}{N_{POS}} \sum_z 1_{\{c_z^* > 0\}} (\mathcal{L}_B \lambda_0 + \mathcal{L}_D \lambda_1) \quad (1)$$

$\mathcal{L}_{GFL}$ is GFL loss scores. N_{POS} represents the number of positive samples. $\mathcal{L}_{QFL}$ indicates Quality Focal Loss. $\mathcal{L}_B$ represents GIoU and $\mathcal{L}_D$ represents Distribution Focal Loss.

NOVELTY AND ALGORITHM

The Novelty of our work includes three parts. First, we carefully selected indiscernible samples to train and adopted a simple heuristic fine-tuning to achieve transfer learning on our dataset. Second, we optimize the hyper-parameter on a target task by fine-tuning, and deploy the program in a real-life scenario to identify the special object. Third, we improve the evaluation of Loss and propose an enhanced algorithm of Generalized Intersection over Union (GIoU) loss(Xu et al.,2021), which can reduce the complexity of time and space.

Table 1 . The algorithm of GIoU loss.

Algorithm 1: Generalized Intersection over Union (GIoU) loss
input: Two arbitrary convex shapes: A, B $\subseteq S \in \mathscr{R}^n$ output: GIoU 1: For A and B, find the smallest enclosing convex object C, where C $\subseteq S \in \mathscr{R}^n$ 2: IoU = $\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$

3: GIoU = IoU- $\frac{|C\backslash(A\cup B)|}{|C|}$

The smallest Closure is a mathematical operation in a set. These elements of the set use all of their relationships to find other elements, with which they have that relationship. The found elements are incorporated into the set until no new element can add into the set. The algorithm of Generalized Intersection over Union (GIoU) loss, as shown in Table 1 , use this operation of exhaustion to absorb now boxes. However, this method could generate some useless computing because of the process of snowballing(Steno et al.,2020). In order to avoid wasting computer resources, we propose an enhanced algorithm, as shown in Table 2.

Table 2 . The algorithm of enhanced GIoU loss.

Algorithm 2: An enhanced GIoU loss algorithm

input: Two arbitrary convex shapes: A, B $\subseteq S \in \mathscr{R}^n$

output: GIoU

1: For A and B, find the smallest enclosing convex object C, where $C \subseteq S \in \mathscr{R}^n$

2: $\exists\ p(x_1, y_1) \in A$, $\exists\ q(x_2, y_2) \in B$, and $p(x_1, y_1)=\max(A)$, $q(x_2, y_2)=\min(A)$

3: $I= A \cap B$, $U = A \cup B$

4: Calculating the area of I:

$$I=\begin{cases} S = L^2|cos\alpha * sin\alpha|, & if\ A \cap B \neq \emptyset . \\ 0, & otherwise. \end{cases}$$

L is a distance from p to q.

5: Calculating the area of U:

U=A+B-I

6: IoU = $\frac{I}{U}$

7: GIoU = IoU - $\frac{|I-U|}{I}$

EXPERIMENTS AND ANALYSIS

We conduct our experiments, respectively, on our dataset and the challenging MS COCO using our enhanced Loss and GFL for object detection. NanoDet YOLO SSD Fast R-CNN and our SAT models are trained on COCO train2017 and our dataset.

Experimental environment: operation system (OS) : Win 10; Development platform: Python3.8+OpenCV+PyCharm+torch1.12.0+cuda116; CPU: Intel(R) Core (TM) i7; GPU: GeForce RTX 3080 SUPER; Memory: 16 G; Disk: 1T.

4.1 Dataset preparation

The dataset of COCO train2017 contains 10000 train images and 5000 validating images with 80 classes, and our dataset contains 1000 train images and 600 validating images, as shown in

Figure 4, which download from the network. More importantly, we annotate a few carefully selected indiscernible samples to strengthen the robustness(Giveki, D.,2021) of models in the training stage.

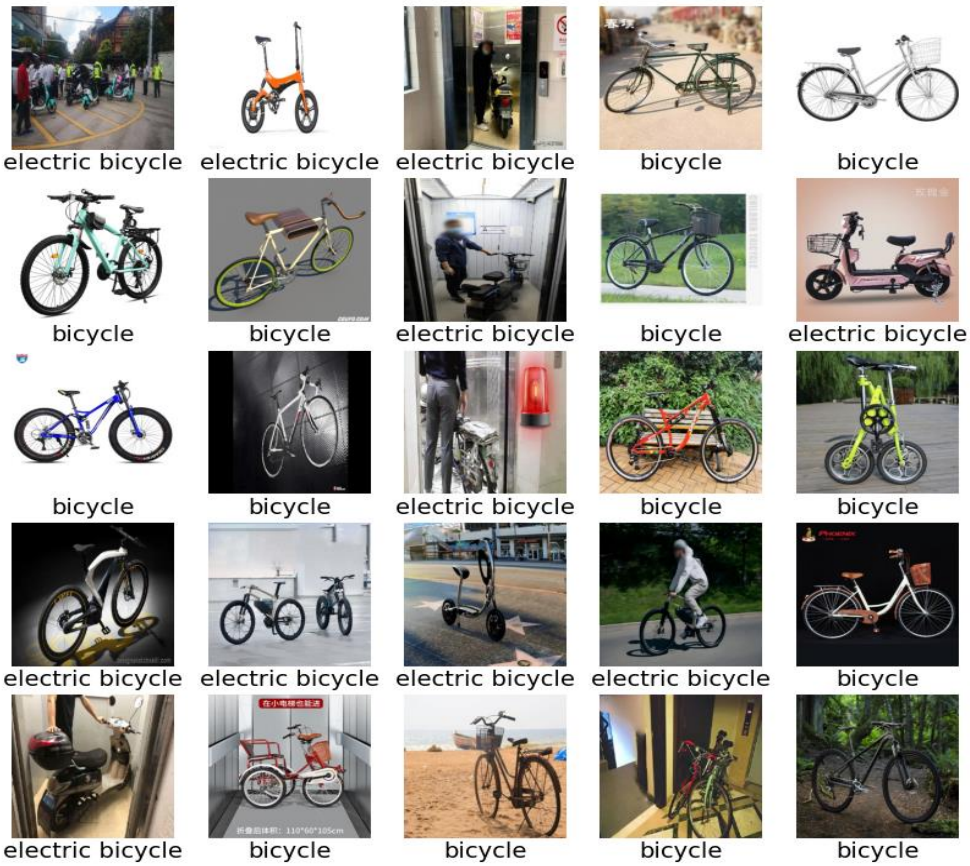

Figure 4 . The images of a part dataset.
(Images show different samples on different occasions)

4.2 Fine-tuning Hyperparameters

All models use uniform batch, normalization and epoch weight normalization(Yan et al.,2020) in the training processes. Hyperparameters(Hou et al.,2021) are configured according to

Table 3, for example, the initial learning rate is 0.14 and the momentum is 0.9. The images are pretreated for 320 × 320 image size by using image cropping and random horizontal mirror scaling. We train NanoDet, YOLO and SAT models for 300 total steps and 190 total epochs to focus on observing epochs and learning rate decay changes at 40, 55, 60, and 65.

Table 3 . The Hyperparameters are configured for SAT in training according to the following table**.**

Hyperparameter	lr	Momentum	Batch	HSB	Total steps	Total epochs

Value	[0.14, 0.1, 0.07,0.01]	[0.9,0.09, 0.0089]	16	[130,160, 175,185]	300	190

For all tasks of identifying bicycle and electric bicycle, we use SGD optimizer(He et al.,2021) with an initial learning rate (lr)(Guo et al.,2022) 0.14, momentum 0.9, and batch size 16. We change input images with small size 98 × 98, medium size 196 × 196 and large size 320 × 320. We adopt a simple heuristic fine-tuning strategy. The configuration of hyperparameter schedule boundaries (HSB) is [130,160,175,185]. Milestones of fine-tuning boost program performance, which make SAT model achieve perfect loss curves on our dataset, as shown in Figure 5-6.

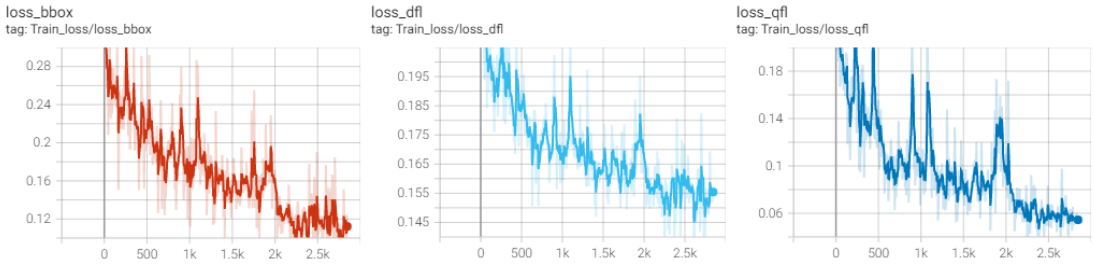

Figure 5 . The courses of loss in 3000 iterations.

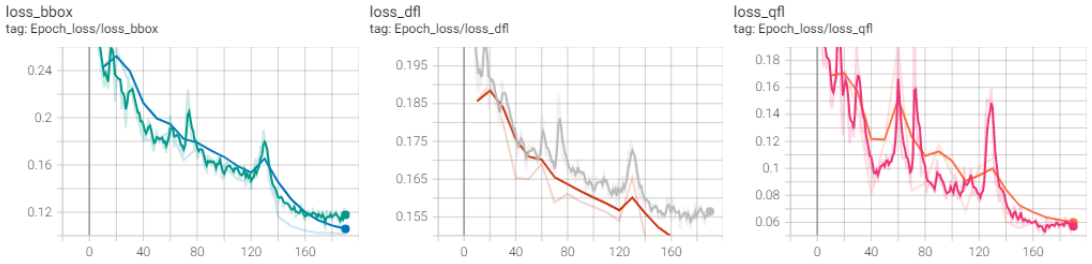

Figure 6 . The courses of loss in 190 epochs.

4.3 Object Detection

At last, we evaluate SAT on object detection and identify images of bicycle and electric bicycle. According to our training protocol, SAT model appears perfect performance, such as mAP(Albuquerque et al.,2021) and maxDets(Shermin et al.,2021). The Average Precision (AP) reaches 95%, the small area metric reaches 90%, the medium area metric reaches 94.8% and the large area metric reaches 94.5%, as shown in Table 4.

Table 4 . The performance of SAT on a special dataset.

Area	mAP	IoU	maxDets		
			1	10	100
all	0.95	0.75	0.796	0.913	0.927
small	0.9	0.5	0.072	0.091	0.095

medium	0.948	0.5	0.169	0.178	0.189
large	0.945	0.5	0.788	0.908	0.946

The maxDets means to retain the first, top 10, and top 100 prediction boxes respectively on each image. Compare these prediction boxes with the real box to calculate the AP and AR. If retain more prediction boxes, it is clear from Table 4 that the mAP of the model would be enhanced. This perfect performance of the model also demonstrates on increasing epochs, as shown in **Figure 7**.

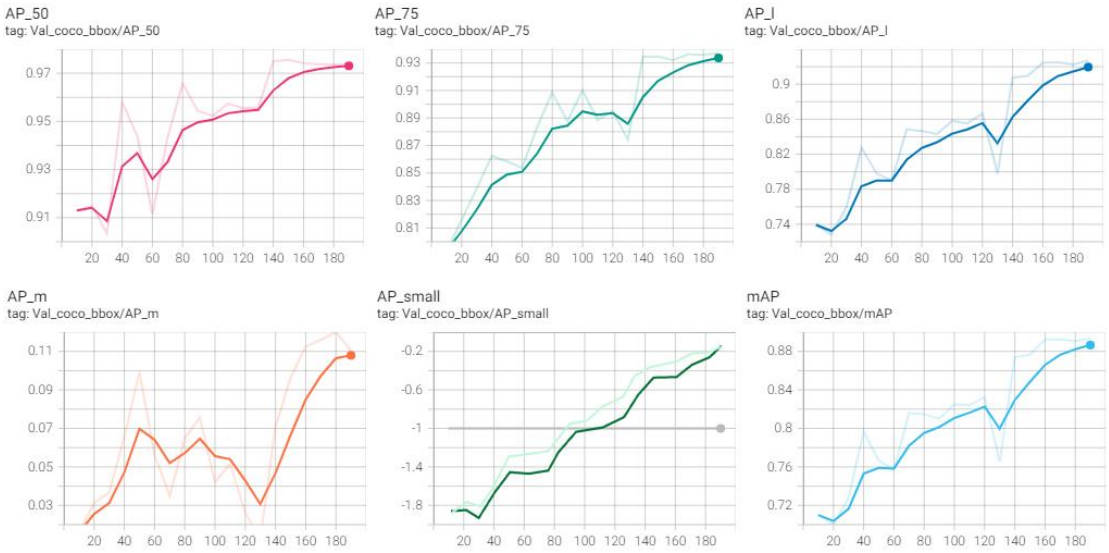

Figure 7 . The performance of SAT with increasing epochs.

We select 9 images with different sizes from real life to verify the accuracy of SAT. The results of the test indicate that our SAT model is perfect to distinguish between bicycle and electric bicycle, which reach 95% AP. Visualization of predicted boxes of the image are drawn using SAT model, as shown in Figures 8-9. Predicted boxes of classification score are also drawn on the picture.

Figure 8 . Group 1: The visualization of predicted boxes of images are drawn by means of SAT.

Figure 9 . Group 2: The visualization of predicted boxes of images are drawn by means of SAT.

COMPARE AND DISCUSSING

Due to memory constraints, we use light model architecture and dataset for all of our models. NanoDet is a light object detection model, which size is less than 4.36M and the amount of parameters is 0.95M, and is super fast on the mobile terminal (10.23ms on ARM CPU). GPU memory cost is much lower than other models: the batch-size 80 on the GTX1060 6G. We

fine-tune NanoDet models based on the strategy of transfer learning to detect special objects. Compare NanoDet, YoloV3 and YoloV4 with our SAT on the latency of processing data, FLOPS (floating-point operations per second)(Gu et al.,2021), Params (the amount of parameters) and model size. The difference of models are shown in Table 5.

Table 5 . Performance comparison of NanoDet, YoloV3, YoloV4 and SAT**.**

Model	Image size	Latency	FLOPS	Params	Model Size
NanoDet	320 × 320	10.23ms	0.72B	0.95M	4.36 mb
YoloV3	416 × 416	37.6ms	5.62B	8.86M	33.7mb
YoloV4	416 × 416	32.81ms	6.96B	6.06M	23.0mb
Our SAT	320 × 320	8.52ms	0.65B	0.84M	4.21mb

In order to indicate the performance of SAT, we compare the speed of different image sizes in the processing of forwarding time, decode time and visualization time, as shown in Table 6.

Table 6. Cost time of three processing stages (Forward, Decode and Visualization).

Image Size	Forward time(s)	Decode time(s)	Viz time(s)
2.1M	2.562	0.016	0.062
1.5M	2.484	0.016	0.063
586K	0.078	0.016	0.016
384K	0.047	0.016	0.031
234K	0.045	0.016	0.016
207K	0.045	0.016	0.016
180K	0.031	0.016	0.016
177K	0.031	0.016	0.016
65K	0.016	0.016	0.016

The trends of the respective time in Forward, Decode and Visualization(Huang et al.,2021) were displayed clearly, as shown in Figure 10. It indicates that processing time decrease as image size reduces. Especially, at the 586K critical point the decode time change largely.

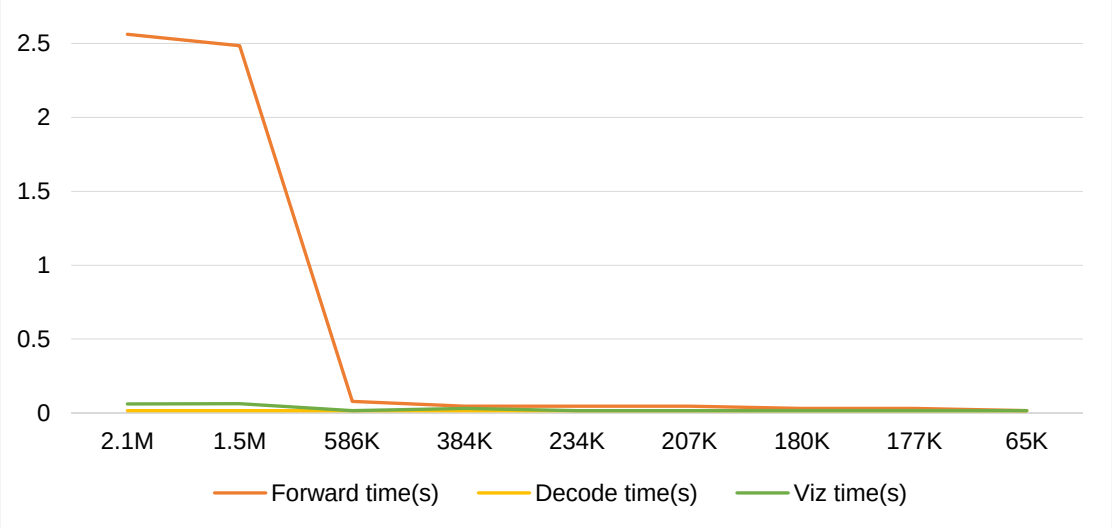

Figure 10 . The trends of cost time (Forward, Decode and Visualization).

CONCLUSION

In computer vision, our SAT modern is a one-stage object detection method. We combine a few carefully selected samples in training and simple heuristic fine-tuning to achieve good performance on special object detection in real life scenarios. SAT performs well across surprisingly small, medium, and large critical areas and achieves 95% AP (Average Precision). Actually, our SAT has the potential to get benefit from NanoDet to further enhance its performance, such as optimizing loss, and deformable network architecture. Due to memory constraints, we only use light model architecture and a partial COCO dataset for training. We hope our work could promote and complete the practical application in more real life scenarios.

ACKNOWLEDGEMENTS

This work is supported by National Nature Science Foundation of China (No. 62262006, No. 12272354), Zhejiang Lab (NO. 2021KE0AB01), Open Fund of Key Laboratory of Monitoring, Evaluation and Early Warning of Territorial Spatial Planning Implementation, Ministry of Natural Resources (No: LMEE-KF2021008), Technology Innovation and Application Development Key Project of Chongqing (No. cstc2021jscx-gksbX0058), Guangxi Key Laboratory of Trusted Software (No. kx202006).

REFERENCES

Piórkowska-Kurpas, A., Hou, S., Biesiada, M., Ding, X., Cao, S., Fan, X., Kawamura, S. & Zhu, Z.-H. 2021. Inspiraling Double Compact Object Detection and Lensing Rate: Forecast for DECIGO and B-DECIGO. The Astrophysical Journal. 908(2) 196-208.

Qi, G., Zhang, Y., Wang, K., Mazur, N., Liu, Y. & Malaviya, D. 2022. Small Object Detection Method Based on Adaptive Spatial Parallel Convolution and Fast Multi-Scale

Fusion. Remote Sensing. 14(2) 420-435.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S. & Houlsby, N. 2019. Big Transfer (BiT): General Visual Representation Learning. arXiv:1912.11370.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z. & Wang, C. 2020. Sparse R-CNN: End-to-End Object Detection with Learnable Proposals. arXiv e-prints. arXiv:2011.12450 1-10.

Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

Junos, M.H., Mohd Khairuddin, A.S., Thannirmalai, S. & Dahari, M. 2021. An optimized YOLO-based object detection model for crop harvesting system. IET Image Processing. 15(9) 2112-2125.

Zhao, L., Zhi, L., Zhao, C. & Zheng, W. 2022. Fire-YOLO: A Small Target Object Detection Method for Fire Inspection. Sustainability. 14(9) 1-14.

Tk, A., Kh, A. & Hb, A. 2020. The Feature Generator of Hard Negative Samples for Fine-Grained Image Recognition. Neurocomputing. 439 374-382.

Wu, X., Sahoo, D. & Hoi, S. 2020. Meta-RCNN: Meta Learning for Few-Shot Object Detection. Association for Computing Machinery. 1679–1687.

Avola, D., Cinque, L., Diko, A., Fagioli, A., Foresti, G. L., Mecca, A., Pannone, D. & Piciarelli, C. 2021. MS-Faster R-CNN: Multi-Stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images. Remote Sensing. 13(9) 1670-1688.

Wang, Z., Xu, P., Liu, B., Cao, Y., Liu, Z. & Liu, Z. 2021. Hyperspectral imaging for underwater object detection. Sensor Review. 41 176-191.

Zhou, Z., Hu, Y., Deng, X., Huang, D. & Lin, Y. 2021. Fault Detection of Train Height Valve Based on Nanodet-Resnet101. 2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC). 709-714.

Tychsen-Smith, L. & Petersson, L. 2017. DeNet: Scalable Real-Time Object Detection with Directed Sparse Sampling. 2017 IEEE International Conference on Computer Vision (ICCV). 428-436.

Chen, Y., Bai, Y., Zhang, W. & Mei, T. 2019. Destruction and Construction Learning for Fine-Grained Image Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5152-5161.

Roh, B., Shin, J. W., Shin, W. & Kim, S. 2021. Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity. Computer Science .abs/2111.14330 1-23.

Zhao, L., Huang, Z. & Wang, W. 2021. A Moving Object Detection Method Using Deep Learning-Based Wireless Sensor Networks. Complexity. 2021 1-12.

Zilberstein, N., Maya, J. A. & Altieri, A. 2021. A BCS microwave imaging algorithm for object detection and shape reconstruction tested with experimental data. Signal processing. **57** 88-91.

Wang, H., Liao, S. & Shao, L. 2021. AFAN: Augmented Feature Alignment Network for Cross-Domain Object Detection. IEEE Transactions on Image Processing. 30 4046-4056.

Xiying, L., Fengwei, Q., Qianyin, J. & Qiang, L. 2022. Structure-guided attention network for fine-grained vehicle model recognition. Journal of Electronic Imaging. 31(2) 023033

Huang, H., Wu, Z., Li, W., Huo, J. & Gao, Y. 2021. Local descriptor-based multi-prototype

network for few-shot Learning. Pattern Recognition. 116 107935-107944.

Zhang, X., Shen, M., Li, X. & Feng, F. 2022. A deformable CNN-based triplet model for fine-grained sketch-based image retrieval. Pattern Recognition. 125 108508-108522.

Xu, K., Li, S., Jiang, X., Lu, J., Yu, T. & Li, R. 2021. A novel transfer diagnosis method under unbalanced sample based on discrete-peak joint attention enhancement mechanism. Knowledge-Based Systems. 212 106645-106660.

Steno, P., Alsadoon, A., Prasad, P. W. C., Al-Dala'in, T. & Alsadoon, O. H. 2020. A novel enhanced region proposal network and modified loss function: threat object detection in secure screening using deep learning. The Journal of Supercomputing. **77** 3840-3869.

Giveki, D. 2021. Robust moving object detection based on fusing Atanassov's Intuitionistic 3D Fuzzy Histon Roughness Index and texture features. International Journal of Approximate Reasoning. 135 1-20.

Yan, T., Wang, S., Wang, Z., Li, H. & Luo, Z. 2020. Progressive Learning for Weakly Supervised Fine-grained Classification. Signal Processing. 171 107519-107528.

Hou, J.-B., Zhu, X. & Yin, X.-C. 2021. Self-Adaptive Aspect Ratio Anchor for Oriented Object Detection in Remote Sensing Images. Remote Sensing. 13(7) 1318-1336.

He, G., Li, F., Wang, Q., Bai, Z. & Xu, Y. 2021. A hierarchical sampling based triplet network for fine-grained image classification. Pattern Recognition. 115 107889-107901.

Guo, C., Lin, Y., Chen, S., Zeng, Z., Shao, M. & Li, S. 2022. From the whole to detail: Progressively sampling discriminative parts for fine-grained recognition. Knowledge-Based Systems. 235 107651-107663.

Albuquerque, C., Vanneschi, L., Henriques, R., Castelli, M., Povoa, V., Fior, R. & Papanikolaou, N. 2021. Object detection for automatic cancer cell counting in zebrafish xenografts. PLOS ONE. 16 e0260609.

Shermin, T., Teng, S. W., Sohel, F., Murshed, M. & Lu, G. 2021. Integrated Generalized Zero-Shot Learning for Fine-Grained Classification. Pattern Recognition.122 108246-108270.

Gu, B., Ge, R., Chen, Y., Luo, L. & Coatrieux, G. 2021. Automatic and Robust Object Detection in X-Ray Baggage Inspection Using Deep Convolutional Neural Networks. IEEE Transactions on Industrial Electronics. 68(10) 10248-10257.

Huang, X., Xu, K., Huang, C., Wang, C. & Qin, K. 2021. Multiple Instance Learning Convolutional Neural Networks for Fine-Grained Aircraft Recognition. Remote Sensing 13(24) 5132-5154.