ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

Study of YOLOX Target Detection Method Based on Stand-Alone Self-Attention

Yanyang Zeng[1,a], Zihan Zhou[1,b,*], Yang Yu[1,c]

[1]College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo City, China
[a]zyyhost@qq.com, [b]isotionismrc@163.com, [c]1139601531@qq.com
[*]Corresponding author

Abstract: *A target detection method based on an improved network of stand-alone self-attention mechanisms (YOLOX_SASA) is proposed to address the problems of complex picture backgrounds, slow detection speed, and low detection accuracy. The method firstly improves the speed of target detection by introducing the stand-alone self-attention module in the multi-scale feature fusion part of YOLOX, so that the network can increase the perceptual field while aggregating the neighborhood information. Secondly, by changing the YOLOX binary classification loss function BCE Loss to MultiLabelMargin Loss for label complementation, which in turn improves the target detection accuracy, and by introducing CutMix data enhancement in the training phase to expand the training set and increase the number of samples. Finally, to test the detection effectiveness of the algorithm, simulation experiments are conducted on a homemade small garbage classification dataset and the PASCAL VOC 2007 public dataset. The experimental results show that the method achieves an average accuracy of 93.81% based on satisfying the real-time performance, which is 4.53% better than the original YOLOX algorithm.*

Keywords: *YOLOX; Stand-alone Self-attention; Target Detection; Multi-scale Feature Fusion; Deep Learning*

1. Introduction

In the field of computer vision, target detection, as one of the most important steps in vision tasks, has a direct impact on the results of subsequent tasks in terms of its performance. Target detection[1] achieves the recognition and localization of objects in images, and its results have important practical value for target tracking[2], action recognition[3], instance segmentation[4], and so on. With the rapid development of artificial intelligence technology and the computing power of hardware devices, deep learning-based target detection methods have gradually become a research hotspot and are widely used in video surveillance, medical diagnosis, autonomous driving, and other fields[5,6].

Early methods of target detection mainly extracted target features manually[7], which had problems such as low detection accuracy and poor applicability. With the development of deep learning, methods to improve target detection performance using convolutional neural networks have received widespread attention. R-CNN[8], as a representative two-stage deep learning target detection algorithm, uses convolutional neural networks to extract features, which effectively improves detection performance, but there is a large amount of redundant computation. Fast R-CNN[9] improves on the former, with the extraction of features from the whole picture feature extraction instead of feature extraction of candidate frames to reduce the computation, and the detection speed is improving while the detection accuracy is improving, but there is still a gap compared with the detection speed of single-stage target detection algorithms. Typical algorithms for single-stage target detection include SSD (Single Shot Multibox Detector) and YOLO series. SSD[10] uses feature maps of different scales to do detection, sets prior boxes with different aspect ratios, and extracts detection results directly using convolution to meet accuracy requirements, but is not very real-time. Unlike SSD, the YOLO series[11-15] generates a priori frames adaptively according to the shape of the real target during the training process, which strikes a balance between target detection accuracy and detection speed. Taken together, the YOLO series algorithms are more capable of meeting the requirements of target detection accuracy and real-time performance compared to the R-CNN, Fast-RCNN, and SSD algorithms.

Among the YOLO series, the YOLOX[16] algorithm is the most up-to-date with the most superior features, but it also has performance that can be improved, based on which many scholars have further improved the accuracy and real-time performance of detection by improving the YOLOX algorithm.

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

Based on the attention mechanism approach, Jianfe Zhang, and Sai Ke[17] enhanced feature representation by adding a lightweight attention module on the CSPLayer layer to make the network pay attention spatially; YongShang L et al[18] improved the feature representation by introducing a spatial attention and channel attention CBAM mechanism (Convolutional Block Attention Module), which improves the correlation of features on space and channels and is more conducive to extracting effective features of the target. Based on the feature extraction network approach, Dengfeng Li[19] designed a lightweight feature extraction network Ghost_ECA, which lightens the model and improves the feature extraction capability. Based on the loss function approach, Huanxin Cheng et al[20] improved the loss function to GIOU[21], which not only considered the overlapping regions differently but also considered the non-overlapping regions to better reflect the overlap and make the prediction frame move toward the target frame. The improvement of the above algorithm increases the depth of the network model, which in turn improves the detection accuracy, but the detection speed decreases and it is difficult to meet the real-time requirements. Compared with channel attention and spatial attention, independent self-attention does not focus on the whole feature map and different location parameters are shared, so its number of parameters is significantly less than that of the convolutional layer, which can improve the detection speed while improving the detection accuracy.

In summary, this paper makes improvements to YOLOX and proposes a YOLOX target detection algorithm based on the independent self-attention mechanism to improve the target detection performance. Firstly, the SASA [22] (Stand-Alone Self-Attention) module is introduced in the YOLOX architecture to make it adaptively focus more on the pixels present in small targets during the down sampling process, and obtain stronger feature representation capability and feature abstraction capability by expanding the perceptual field to improve the detection capability and detection speed; secondly, the binary classification loss function is changed Finally, in order to enrich the data set and add small target samples, CutMix is used for data enhancement. By analyzing the simulation results, we can obtain that the accuracy of the YOLOX_SASA target detection method improves by 5.15% over the original YOLOX and 4.6% over the YOLOX_cbam method in the homemade data set, and the accuracy of YOLOX_SASA target detection method improves 3.91% over the original YOLOX and 2.78% over the YOLOX_cbam in the public data set.

2. YOLOX System Model Based on Stand-Alone Self-Attention Mechanism

We only accept papers wYOLOX can be divided into three modules, which are the feature extraction network module (Backbone), the multi-scale feature fusion module (Neck), and the result output module (Output). The structure of the YOLOX_SASA network proposed in this paper is shown in Figure 1.

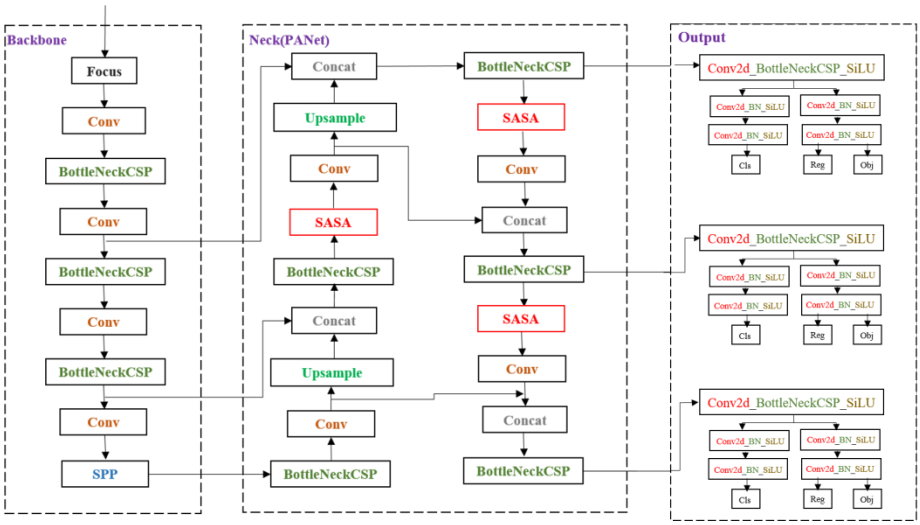


Figure 1: The structure of improved YOLOX_SASA network.

The workflow of the YOLOX network is as follows. Step 1: The image enters from the input to the Backbone network model for feature extraction, and the extracted features are feature layers, and three effective feature layers are obtained in the backbone part for the next step of network construction. Step 2: The three effective feature layers obtained in the backbone network part are fused in FPN (Feature Pyramid Network) for multi-scale fusion, which aims to combine feature information from different scales, and in the FPN part, the effective features that have been obtained are used to continue feature

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

extraction. After the PAN (Path Aggregation Network) structure, the network not only upsamples the features but also downsamples them again to achieve feature fusion. After downsampling, the network feature abstraction capability is improved by expanding the perceptual field with stand-alone self-attentive models. Step 3: The Output layer is mainly the classifier and regressor of YOLOX. At this time, the feature map obtained by the backbone network and Neck network can be regarded as a collection of multiple feature points, so the work done by the Output layer is to judge the feature points and determine whether there are objects corresponding to them. Finally, after the score screening and non-maximum suppression, the prediction frame with the score satisfying the confidence level and the prediction frame with the largest score belonging to the same species in a certain region are screened, and the final detection result is output.

3. Solutions to Enhance the Effectiveness of YOLOX Detection

The YOLOX network is mainly used for general-purpose target detection. To improve the accuracy of target detection, the system model is improved by adding the stand-alone self-attention module, data enhancement, and changing the binary loss function. The number of parameters is reduced by replacing the convolution with an independent self-attentive module to meet the real-time requirement and improve the target detection accuracy at the same time.

3.1. Introduction of Stand-Alone Self-Attention Module

The structure of the spatial self-attention mechanism module is shown in Figure 2. Firstly, the query of the input image and each key are calculated for similarity to get the weights, then these weights are normalized by softmax, and finally, the normalized weights are weighted and summed with the corresponding value to get the final output.

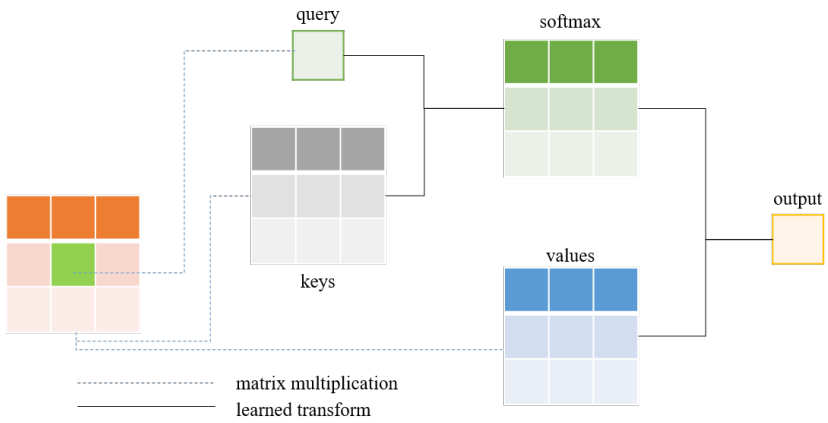


Figure 2: Self-attention over spatial schematic.

The formula of self-attention is shown in Equation (1).

$$Y_{i,j} = \sum_{a,b \in N_k(i,j)} soft\max_{ab}\left(q_{ij}^T k_{ab}\right) v_{ab} \tag{1}$$

Where $q_{ij} = W_Q x_{ij}$, $k_{ij} = W_K x_{ij}$, $v_{ij} = W_V x_{ij}$ and $W_Q, W_K, W_V \in R^{d_{out} \times d_{in}}$ are the learned parameters. However, formula (1) does not contain location information, so for a query, its neighborhood information location relationship cannot be reflected. To address this problem, the position information is added to the self-attentive operation by embedding vectors to represent the relative position, and the improved formula is shown in Equation (2).

$$Y_{i,j} = \sum_{a,b \in N_k(i,j)} soft\max_{ab}\left(q_{ij}^T k_{ab} + q_{ij} r_{a-i,b-j}\right) v_{ab} \tag{2}$$

where q, k, v are the linear projection of the input features, and $r_{a-i,b-j}$ is the relative position embedding of (i, j) and (a, b). From the formula, the transformed central pixel is used as the query, k_{ab} and v_{ab} are summed in the neighborhood. The $Soft\max$ function calculates the weights by the

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

distance between k_{ab} and the query. Thus, instead of convolution, the independent self-attention module achieves weight sharing between different spatial locations, while reducing the number of parameters used in the network and improving the detection speed. The YOLOX_SASA network model after incorporating local self-attention extends the ability to focus on different locations and improves the performance in training and inference.

3.2. Improved Dichotomous Loss Function

The BCE Loss binary cross-entropy loss function used in the YOLOX target detection algorithm for class loss is shown in Equation (3).

$$loss = \frac{1}{N}\sum_{n=1}^{N} l_n \tag{3}$$

Where $l_n = -w\left[y_n \times \log x_n + (1-y_n) \times \log(1-x_n)\right]$ is the loss corresponding to the nth sample, w is the hyper-parameter, y is the true class, and each element takes the value of 0 or 1. That is, it corresponds to the batch data D(x,y) containing N samples, and one input sample corresponds to one classification output.

For the case that one image in the data set of this article contains multiple classifications, the binary classification loss function is improved to multi-label classification, and the improved loss function is shown in Equation (4).

$$loss = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M}\sum_{i=1}^{M} l_n^i \tag{4}$$

where $l_n^i = -w_i\left[y_n^i \times \log x_n^i + (1-y_n^i) \times \log(1-x_n^i)\right]$, w_i is the hyper-parameter, which is used to deal with the sample imbalance between labels. The improved multi-label classification loss function takes the method of label complementation when the number of labels of the image is not fixed, and the missing labels are labeled using 0 to improve the target detection performance and reduce the case of missing detection.

3.3. Data Enhancement Strategy

The model uses CutMix data augmentation to increase the number of positive and negative samples to achieve a balance and enrich the dataset to address the problem of an unbalanced number of categories in target detection.

After data augmentation, the pixels in the training set are all information pixels in the training set, thus improving the training efficiency; at the same time, the model can further enhance the localization ability of the model by locally identifying objects and adding information of other samples to the cropped regions. By using the CutMix data enhancement method, the performance of the model classification is improved while the training and inference costs remain unchanged.

The results after using the CutMix data enhancement method on the dataset PASCAL VOC 2007[23] are shown in Figure (3).

Figure 3: CutMix Data Enhancement.

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

4. Experimental Numerical Simulation Analysis

4.1. Datasets

The datasets in this paper are selected from the self-made garbage classification Self-made_garbage data set and the publicly available PASCAL VOC 2007 data set. Most of the sources of the self-made data set are small target images filtered from the crawler results, and a few of them are real images. To verify the effectiveness of the proposed method, this paper also uses the publicly available PASCAL VOC 2007 data set for testing. The distribution of the number of training set test sets included in the experimental data set is shown in Table 1.

Table 1: This caption has one line so it is centered.

Datasets	Test set	Training set
Self-made_garbage	7453	5598
PASCAL VOC 2007	5011	4952

There are 40 categories of self-made_garbage. In this paper, the images are firstly annotated in YOLO format, and after the annotation is completed, the txt files corresponding to the images are downloaded and classified into the training set and test set. There are 20 categories in the PASCAL VOC 2007 data set, among which there are 5011 images in the training set and 4952 images in the test set, totaling 9963 images.

4.2. Experimental Configuration

The experimental environment in this paper uses Windows 10 operating system, Tesla K80 for computing, PyTorch version 1.9.0, and Python version 3.9. The initial learning rate in this paper is set to 0.0032, the total number of iterations is 300, and the iteration batch size is set to 32.

The experimental configuration is shown in Table 2.

Table 2: Experimental configuration.

Serial number	Configuration	*Parameter*
1	GPU	Tesla K80
2	Deep learning framework	PyTorch 1.9.0
3	Development Environment	Python3.9
4	Batch size	32
5	Number of training rounds	300
6	Initial learning rate	0.0032

4.3. Evaluation Metrics

The evaluation metrics in this paper use the average accuracy mean(*mean Average Precision* , *mAP*) and average precision(*Average Precision* , *AP*) to evaluate the performance of the target detection method, and the more their values converge to 1, the better the target recognition effect. The *mAP* and *AP* need to be calculated by the accuracy *Precision* and recall *Recall* of the model training samples, whose expressions are shown in Equation (5).

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \end{aligned} \tag{5}$$

Where *TP* is the number of positive samples correctly identified, *FP* is the number of negative samples identified as positive, and *FN* is the number of positive samples identified as negative. The average precision (*AP*) of each target category can be calculated from the area under the accuracy-recall curve (*P* - *R* curve) composed of *Precision* and *Recall* and the formula is as in Equation (6).

$$AP = \int_0^1 P \mathrm{d}R \tag{6}$$

The average accuracy (*mAP*), which is the average accuracy summed over all categories and divided by the total number of categories, is calculated as in Equation (7).

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

$$mAP=\frac{1}{n}\sum_{i=0}^{n}AP_i \tag{7}$$

n is the total number of categories in the training data set, and i is the number of the current category.

4.4. Ablation Experiments

To verify the effectiveness of the stand-alone self-attention mechanism, CutMix data enhancement, and multi-categorization loss function improvement strategy on the improvement of model detection, YOLOX was used as the benchmark for ablation experiments, and the same experimental equipment, as well as data, were used for testing. mAP and inference time per frame was selected as the experimental evaluation indexes, "×" indicates that the improvement strategy was not used, and "√" indicates that the improvement strategy was used. the experimental results of each model are shown in Table 3.

Table 3: Ablation experiment results.

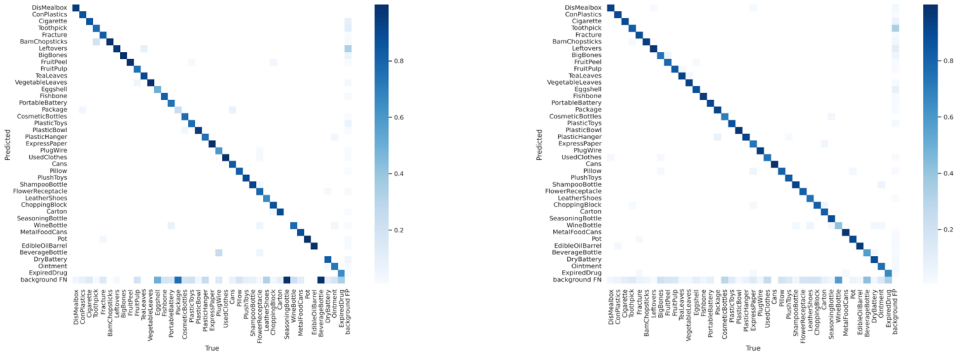
Improvement Name	Stand-Alone Self-Attention	*Data Enhancement*	Multi-Categorization Loss	*mAP*	*Inference Time per Frame/ms*
YOLOX	×	×	×	0.8903	35
Improvement 1	√	×	×	0.9259	32
Improvement 2	×	√	×	0.9134	35
Improvement 3	×	×	√	0.9191	35
Improvement 4	√	√	×	0.9314	32
Improvement 5	√	×	√	0.9397	33
Improvement 6	√	√	√	0.9419	32

Comparing all models of ablation experiments, after gradually adding modules to the original YOLOX model, it can be seen that adding modules in the experiments improves the average detection accuracy of the model, the inference time per frame is reduced, and the detection speed of the improved model still meets the real-time detection requirements. The improved YOLOX algorithm in this paper has the highest detection accuracy and maintains a good real-time performance, and the overall performance is outstanding, and the improved algorithm has obvious superiority compared with other algorithms.

4.5. Analysis of Simulation Experiment Results

To verify the effectiveness of this algorithm, training and testing were performed on the homemade data set Self-made_garbage, and the original YOLOX network was chosen for comparison tests.

A confusion matrix is a form of evaluation of accuracy in the classification problem. The classification effectiveness of the model is evaluated by observing the diagonal of the confusion matrix, and the highest classification accuracy is indicated when all the data are on the diagonal. The results of the confusion matrix are shown in Figure 4. From the figure, it can be seen that the data of YOLOX_SASA are all on the diagonal, which is significantly better than the YOLOX network.

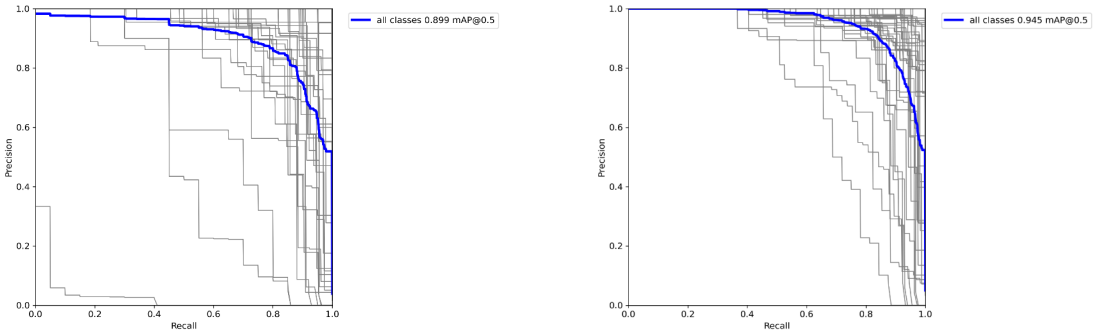


(a) YOLOX (b)YOLOX_SASA

Figure 4: Graph of confusion matrix results.

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

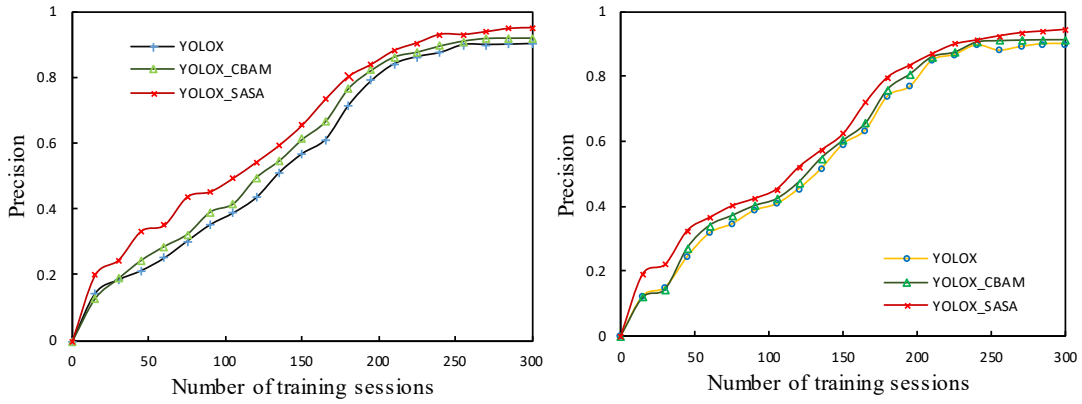
The YOLOX model has an average accuracy mean of 89.9% after 300 rounds of training; the YOLOX_SASA model has an average accuracy mean of 94.5% after 300 rounds of training, and the experimental results are shown in Figure 5. The area enclosed by each gray line and x-axis (Recall) and y-axis (Precision) is the average accuracy of each category, and the analysis of the figure shows that the area enclosed by the YOLOv5s_SASA network model is larger than that enclosed by the YOLOv5s network model, so the detection accuracy of YOLOv5s_SASA is stronger than that of the YOLOv5s target detection algorithm.



(a)YOLOX (b)YOLOX_SASA

Figure 5: mAP for the YOLOX model and YOLOX_SASA.

To further illustrate the effectiveness of the proposed method, a comparison test was done in the public data set PASCAL VOC 2007, and the experimental results are shown in Figure 6. In the homemade data set, the YOLOX_SASA network model has a better recognition effect of 94.54%, which is 4.36% better than the YOLOX_CBAM network model and 4.62% better than the original YOLOX target detection algorithm. In the PASCAL VOC 2007 data set, the YOLOX_SASA network model achieves 93.08%, which is 2.98% better than the original YOLOX_CBAM network model and 4.43% better than the original YOLOX target detection algorithm. It can be seen based on Figure 5 and Figure 6 that the performance of the network model proposed in this paper is improved in both the homemade data set and the public data set tests.



(a)Experimental results of the three algorithms on the Self-made_garbage data set

(b)Experimental results of the three algorithms on the PASCAL VOC 2007 data set

Figure 6: The experimental results in different datasets.

The comparison results of target detection in the homemade dataset are shown in Figure 7. From Figure 7, it can be seen that the target detection method before the improvement has low confidence and incomplete target detection, and the improved target detection method not only improves the confidence of the target, but also frames the position more accurately, and reduces the target leakage detection at the same time. In this paper, the YOLOX_SASA target detection method performs outstandingly, with a significantly higher recognition rate and significantly fewer missed detections.

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

(a)Before improvement (b)After improvement

Figure 7: Comparison of test results.

5. Conclusions

The proposed YOLOX_SASA target detection method optimizes the original YOLOX in 3 aspects: data enhancement, the introduction of a stand-alone self-attention mechanism, and changing the loss function, which effectively improves the cyberspace target localization and solves the problem of weak target feature representation. The validation results on the homemade garbage classification dataset and the PASCAL VOC 2007 public dataset show that the YOLOX_SASA target detection method improves the detection accuracy without any decrease in detection speed. However, there are still areas where the method in this paper can be improved, and subsequent research will further optimize the network to improve the performance of the algorithm comprehensively.

References

[1] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey [J/OL]. (2019-05-16)[2021-11-15]. https://arxiv.org/abs/1905.05055.
[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers", 2020.
[3] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," IEEE Transactions on Image Processing, vol.27, no.3, PP.1347-1360, 2018.
[4] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[5] Xu D G, Wang L, Li F. A review of research on typical target detection algorithms for deep learning [J]. Computer Engineering and Applications, 2021, 57(08):10-25.
[6] Zhao Y-Q, Rao Y-Y, Dong S-P, Zhang J-Y. A review of deep learning target detection methods [J].

ISSN 2616-5775 Vol. 5, Issue 12: 29-37, DOI: 10.25236/AJCIS.2022.051205

Chinese Journal of Graphics, 2020, 25(04):629-654.
[7] Harzallah H, Jurie F, Schmid C. Combining efficient object localization and image classification[C]//2009 IEEE 12th international conference on computer vision. IEEE, 2009: 237-244.
[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[9] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
[10] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
[11] 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. 2016: 779-788.
[12] REDMON J, FARHADI A.YOLO9000: better, faster, stronger[C]/Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017:6517-6525.
[13] REDMON J, FARHADI A.YOLOv3: an incremental improvement[C]/Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:89-95.
[14] BOCHKOVSKIY A. WANG C Y, LIAO H-Y M. YOLOv4: Optimal Speed and Accuracy of object Detection [J]. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[15] Jocher G. Yolov5 [J]. Code repository https://github.com/ultralytics/yolov5, 2020. 019
[16] GE Z, LIU S, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[J]. ArXiv Preprint, 2021. ArXiv: 2107.08430.
[17] Zhang Jianfei, Ke Sai. Research on improving YOLOX fire scene detection method [J]. Computer and Digital Engineering, 2022, 50(02):318-322+349.
[18] Y. S. Li, R. G. Ma, M. Y. Zhang. Improving YOLOv5s+DeepSORT for monitoring video traffic statistics [J]. Computer Engineering and Applications, 2022, 58(05):271-279
[19] Dengfeng Li, Ming Gao, Wentao Ye. A ship target detection algorithm combining lightweight feature extraction network [J]. Computer Engineering and Applications: 2022, 1-10.
[20] Cheng, X. X., Jiang, Z. Q., Cheng, L., Cheng, K. Improved YOLOX-S-based algorithm for helmet reflective clothing detection [J]. Electronic Measurement Technology, 2022, 45(06):130-135.
[21] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalied intersection over union: A metric and a loss for bounding box regression[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.658-666.
[22] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019.
[23] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge [J]. Intemational journal of computer vision, 2010, 88(2): 303-338.