Research on Dynamic Object Detection Algorithm based on Improved YOLOv5s

Haoran Ding[1,2], Sanpeng Deng[1,2], Tianhui Liu[1,2], Chuanqing Ma[1,2]

[1] Robot and Intelligent Equipment Research Institute of Tianjin University of Technology and Education, Tianjin 300222, China

[2] Tianjin Enterprise Key Laboratory of Intelligent Robot Technology and Application, Tianjin 300350, China

Abstract

In complex dynamic environments, YOLOv5s not only consume a lot of resources and time to process irrelevant background messages, but also may cause image distortion and loss of information about targets in dynamic environments, which not only leads to a reduction in the computational efficiency of the network model, but also leads to a prolongation of the inference time, which is not conducive to real-time dynamic object detection. To address these problems, an improved YOLOv5 algorithm is proposed, which introduces a hybrid attention mechanism before the convolution module of the Neck network, in addition to replacing the nearest-neighbor interpolation method originally used for up-sample with an anti-convolution method. It is experimentally verified that the accuracy of the improved YOLOv5s algorithm is 5.75% higher than that of the YOLOv5s algorithm in a dynamic environment.

Keywords

Dynamic Environments; YOLOv5s; Hybrid Attention Mechanisms; Deconvolution Methods.

1. Introduction

Object detection plays an important role in the field of computer vision. It refers to algorithms and techniques to recognize and localize target objects of interest in an image or video. The development of object detection is crucial for the realization of tasks such as image description generation, target tracking, and scene understanding. With the rapid development of artificial intelligence technology and the upgrading of computer hardware and equipment, object detection algorithms have made a breakthrough exhibition. Traditional object detection algorithms mainly rely on hand-designed feature extractors and classifiers, while modern marker detection algorithms, such as deep learning-based algorithms, are able to automatically learn features and perform end-to-end target measurement.

In order to improve the accuracy of object detection in dynamic environments, a large number of related researches have been conducted at home and abroad to address this problem. The three improvements proposed by Jia[1] for YOLOv5s include adding P2 detection head for detecting smaller targets; adding CBAM attention mechanism to Neck network to improve the network's attention to small targets; and optimizing the feature fusion structure of YOLOv5s by borrowing the idea of BiFPN network. Zhao[2] proposed three improvement schemes based on the structure of YOLOv5s, firstly, the multi-scale feature detection is improved, not only the multi-scale range is increased but also a new feature fusion structure is added; secondly, the attention mechanism is added to the model and the channel attention mechanism is improved; finally, a Gaussian filter is introduced in the object detection, which suppresses the Gaussian noise and enhances the feature extraction.

Although the above study improved YOLOv5s, it did not improve the detection accuracy of dynamic targets. In complex dynamic environments, YOLOv5s not only consumes a lot of resources and time to process irrelevant background messages, but also may cause image distortion and loss of information about targets in dynamic environments, which not only leads to a reduction in the computational efficiency of the network model, but also leads to a prolonged inference time, which is not conducive to real-time dynamic object detection. To address the above problems, an improved YOLOv5s network model is proposed, aiming to improve the accuracy of object detection in dynamic environments.

2. YOLOv5s Object Detection Algorithm

2.1 The Structure of the YOLOv5s Network

YOLOv5 creates four YOLOv5 network models with different sizes and complexities of network structures by setting different depth_multiple and width_multiple parameters: the YOLOv5s (Small), YOLOv5m (Medium), YOLOv5l (Large), and YOLOv5x (Xtra Large).YOLOv5s is the smallest model, having the least number of parameters and computational complexity, and is the basis for the other three models[3].The structure of the YOLOv5s network is shown in Fig. 1.

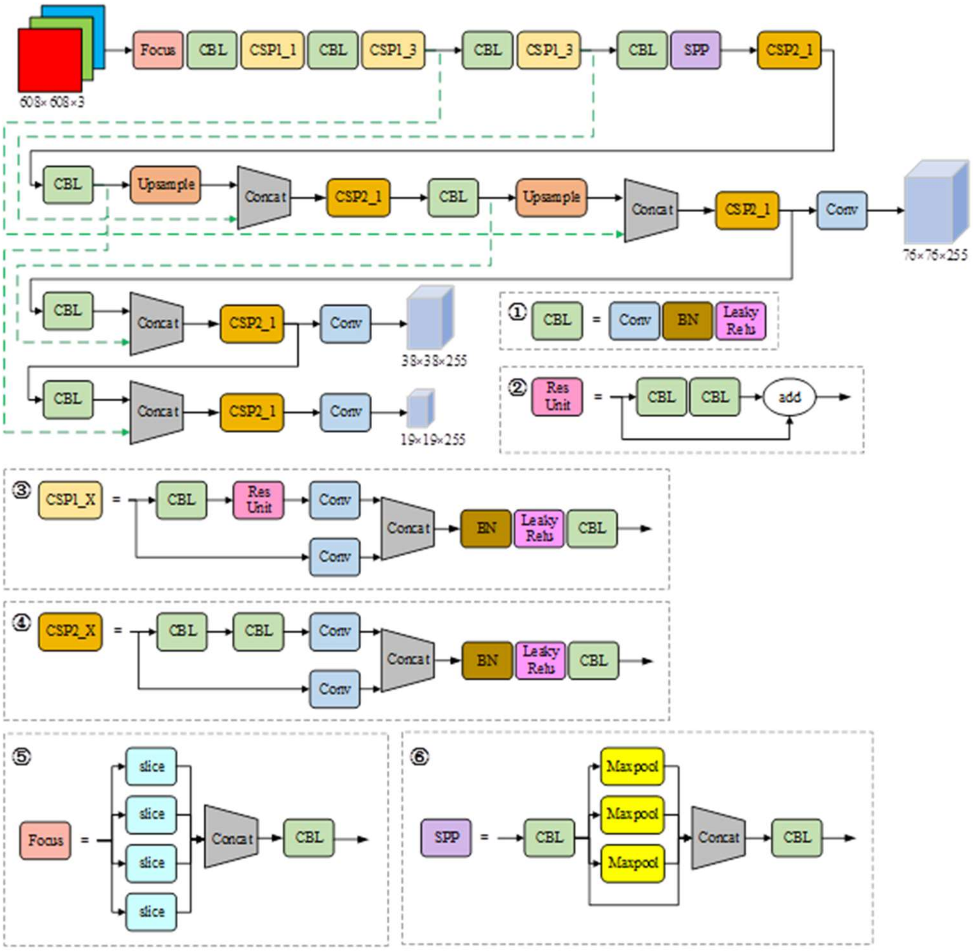

608×608×3
Focus
CBL
CSP1_1
CBL
CSP1_3
CBL
CSP1_3
CBL
SPP
CSP2_1
CBL
Upsample
Concat
CSP2_1
CBL
Upsample
Concat
CSP2_1
Conv
76×76×255
CBL
Concat
CSP2_1
Conv
38×38×255
CBL
Concat
CSP2_1
Conv
19×19×255
① CBL = Conv BN Leaky Relu
② Res Unit = CBL CBL add
③ CSP1_X = CBL Res Unit Conv Conv Concat BN Leaky Relu CBL
④ CSP2_X = CBL CBL Conv Conv Concat BN Leaky Relu CBL
⑤ Focus = slice slice slice slice Concat CBL
⑥ SPP = CBL Maxpool Maxpool Maxpool Concat CBL

Fig. 1 The structure of the YOLOv5s network

The YOLOv5s network structure consists of four main parts: Input, Backbone, Neck and Output.

(1) Input:The YOLOv5s input is used to receive the image and preprocess the image, including Mosaic image enhancement, adaptive image scaling, and adaptive anchor frame computation[4]. The above three operations enable the model to better adapt to target objects of different scales and improve the accuracy, robustness and generalization of detection.

(2) Backbone:In YOLOv5, the Backbone layer refers to the backbone network part of the model, which is usually used to extract the features of the input image.The Backbone layer used in YOLOv5s is based on the CSPDarknet53 architecture, which is inspired by Darknet and CSPNet.The architecture is a network architecture based on residual structures and channel attention mechanisms , which consists of multiple residual blocks and cross-stage connections, and efficiently extracts and represents image features through the stacking and interaction of these structures. Due to the problem of repeated computation of gradients during network optimization, which leads to excessive computation of the inference process[5], YOLOv5s introduces the CSP1_X module in its backbone network structure, which is shown in box ③ in Fig. 1.

(3) Neck:The Neck Network of YOLOv5s mainly consists of two parts, FPN (Feature Pyramid Networks) and PAN (Pyramid Attention Network)[6].FPN is designed to overcome the two problems of insufficient multiscale information and incomplete feature pyramid, which is achieved through the bottom-top connection , side connections and feature fusion operations to gradually extract features and fuse multiscale information from the bottom to the top level. PAN, on the other hand, is designed to effectively fuse features from different scales to improve the detection performance, and it constructs a complete feature pyramid through path aggregation operations and effectively integrates the feature information from different scales.This top-down and bottom-up fusion of FPN+PAN gives full play to the advantages of FPN and PAN, as it can deliver both strong semantic information downward and strong localization information upward.

(4) Output:The output side is also known as Head layer, which is the last layer in the YOLOv5s network model and is used to perform classification and regression operations on the input feature maps to achieve the object detection task. The loss function and NMS play a very important role in the output side.The loss function of YOLOv5s consists of three parts: confidence loss, classification loss and bounding box loss.The confidence loss and classification loss both use the Binary Cross Entropy Loss (BCEWithLogitsLoss) loss function, and the bounding box loss uses the CIoU (Complete IoU) loss function[7] to measure the prediction accuracy of object detection, which can more accurately measure the gap between the predicted bounding box and the true bounding box than the traditional IoU loss function.

2.2 Problems with YOLOv5s Object Detection Algorithm

In complex dynamic environments, YOLOv5s will consume a lot of resources and time to process irrelevant background messages, which will not only lead to a reduction in the computational efficiency of the network model, but also lead to a prolonged inference time, which is not conducive to real-time object detection, so an attention mechanism needs to be introduced to solve this problem. At the same time, the nearest neighbor interpolation method used in the up-sample module may cause image distortion and loss of information about the target in the dynamic environment, so the up-sample module needs to be improved to solve this problem.

3. Object Detection based on Attention Mechanism

3.1 Reasons for Introducing Attention Mechanisms

Attention mechanism is a technique used to enhance the expressive power and accuracy of deep learning models, which helps the model to selectively focus on features relevant to the current task in the input data so that the model can better capture and utilize key information. Common attention mechanisms include Channel Attention SE[8] (Squeeze-and-Excitation), CA (Channel Attention Module), ECA (Efficient Channel Attention Module), Spatial Attention SA (Spatial Attention) and Hybrid Attention CBAM[9] (Convolutional Block Attention Module), etc.

The attention mechanism is introduced into YOLOv5s network structure for three reasons: first, the attention mechanism makes the network more focused on the target object, which helps to improve the detection accuracy and recall of YOLOv5s; second, the attention mechanism helps to extract the important features in the image, suppresses the noise and irrelevant information, and helps to enhance the robustness of YOLOv5s in complex scenes (e.g. shading, changes in light intensity, etc.); third, the attention mechanism guides the network to learn the image features more efficiently, which helps to accelerate the convergence of the model and reduce the resource consumption.

3.2 Improvements based on the Attention Mechanism

In the choice of attention mechanism, because channel attention loses part of the spatial information when embedding channels and spatial attention is difficult to capture global dependencies, a hybrid attention mechanism combining channel attention with spatial attention is chosen. CBAM is an attention module based on the combination of channel attention CA and spatial attention SA, which first adaptively weights each channel by CA, and then use the SA to capture the spatial correlation of the feature map. However, CBAM learns the information of feature maps through global pooling will lose some local information and pay insufficient attention to small-scale objects. Therefore, in this paper, we combine channel attention ECA and spatial attention SA to compose a new hybrid attention mechanism, which is called ECA_SA module in this paper. Compared with CBAM, ECA_SA pays more attention to adaptive weighting in channel and spatial dimensions respectively, adopts separable convolution operation, and performs better in dealing with small-scale objects.

The choice of where to add the attention mechanism is also crucial to the optimization of the model.The Neck network is located between the backbone network and the output layer, and plays a role in the YOLOv5s model, which is responsible for further processing, fusion and aggregation of the features extracted from the backbone network. While in Neck network, adding the attention mechanism before the convolution operation helps the network to better capture the correlation between features. Therefore, in this paper, we choose to add the ECA_SA module before each convolutional block in the Neck network, and the new network model is called YOLOv5s_ECA_SA, whose network structure is shown in Fig. 2. When adding the ECA_SA module, it is necessary to make sure that the number of channels of the ECA_SA module is consistent with that of its preceding and following modules, and at the same time, the connectivity of the ECA_SA module with other modules in the network model should be checked.

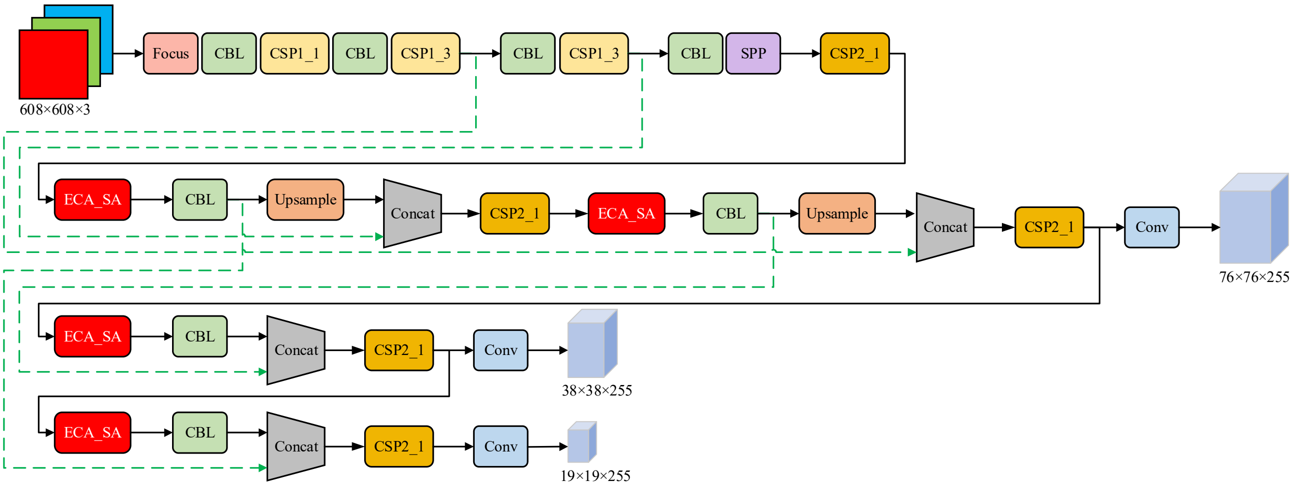


Fig. 2 YOLOv5s_ECA_SA network structure

4. Improved Up-sample Methods

4.1 Reasons for Improving Up-sample Methods

Upsample is an operation that increases a low resolution image or feature map to the same size as a high resolution image or feature map. In YOLOv5s network model up-sampling used in the nearest neighbor interpolation, it is a simple and intuitive up-sampling method, in the up-sampling it simply copies the values of the nearest neighbor pixels, and can not be effective reconstruction of the image texture and details, so the image distortion phenomenon will occur.

4.2 Improved Up-sample Methods

In order to solve the problem of losing image information in dynamic environments, the original nearest-neighbor interpolation will be replaced by an deconvolution method, which is used to back-propagate the gradient by learning the parameters to map the low-resolution feature map back to the high-resolution space. Deconvolution up-sample schematic shown in Fig. 3, the figure illustrates the two convolution methods, the convolution kernel are 3, fill are 0, the difference is that the way A convolution step is 1, the way B convolution step is 2. From Fig. 3, we can see that the deconvolution can be adjusted by adjusting the step size to determine the sampling interval, so as to flexibly control the speed of the up-sample, and the speed is directly proportional to the step size.

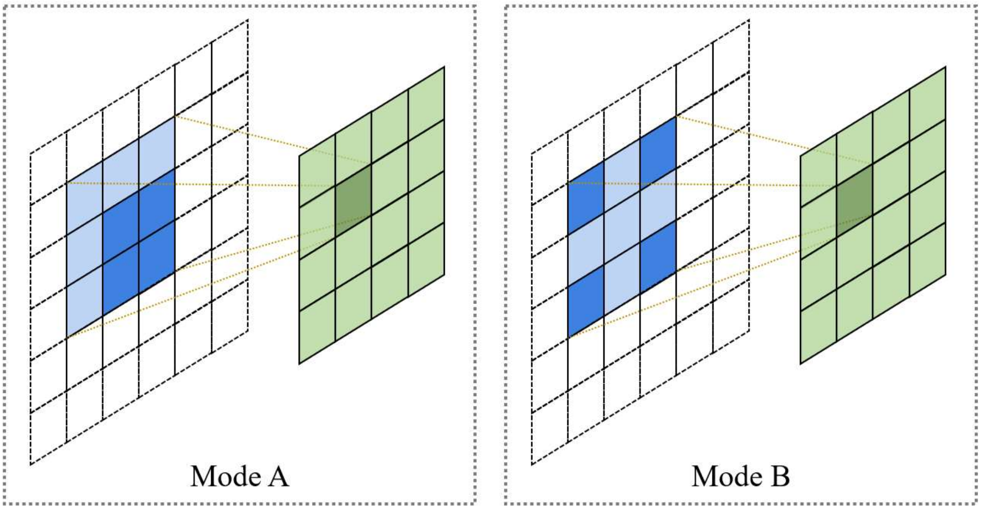


Fig. 3 Schematic of deconvolution up-sample

The kernel tensor is also an important parameter in the deconvolution method. The kernel tensor, also known as the weight matrix, determines the feature map mapping relationship, usually denoted by (I, O, w, h), I is the number of input channels, O is the number of output channels, and w and h are the height and width of the convolution kernel. The input low-resolution feature map is multiplied element by element with the kernel tensor, placed in the corresponding position, and finally the obtained outputs are summed up to obtain the high-resolution feature map, which is shown schematically in Fig. 4.

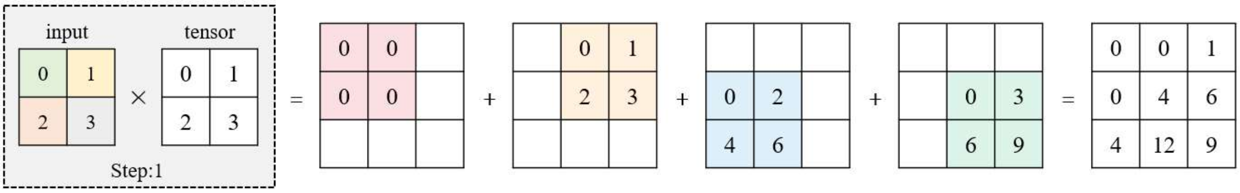


Fig. 4 Schematic diagram of the deconvolution calculation

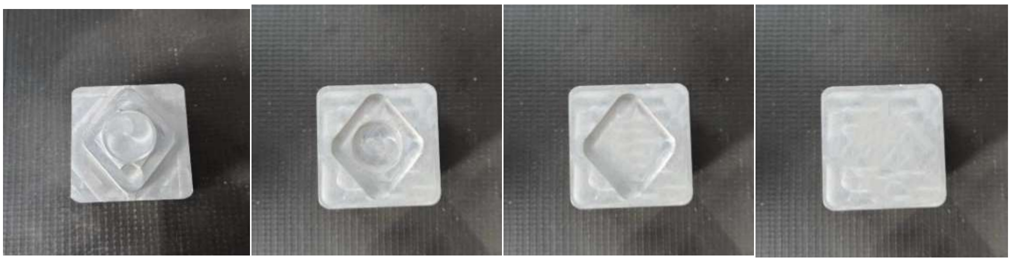
The nearest-neighbor interpolation method in the up-sample module in layers 13 and 18 of the YOLOv5s_ECA_SA network model is replaced with the deconvolution method, and the relevant parameters to be configured for the inverse convolution are shown in Table 1. The replaced network model is called YOLOv5s_ECA_SA_Deconv network model.

Table 1. Inverse Convolution Parameter Configuration

Network layer	Input Number of channels	Output Number of channels	Convolutional Kernel Size	Convolutional step	Input side fill value	Output side fill value	Channel Blocking Connections
13	256	256	4	2	1	0	256
28	128	128	4	2	1	0	128

5. Experiment and Analysis

In order to verify the detection performance of the improved YOLOv5s network model for targets in complex environments, a large amount of data on targets in complex environments were collected as datasets, and training and testing were performed on these datasets. As shown in Fig. 5, the experimental scene is selected as a three-dimensional warehouse, and the targets are selected as 4 kinds of machining blanks.



(a) TOOL_A (b) TOOL_B (c) TOOL_C (d) TOOL_D

Fig. 5 Four types of blanks

Use labelImg software to add labels to two types of blanks, category TOOL_A, TOOL_B, TOOL_C and TOOL_D respectively. convert the labeled dataset to YOLO format data, after the conversion is complete, the labeled data is randomly assigned to the training set and validation set according to the ratio of 8:2, i.e., randomly select 80% (480 pictures) of the images in the dataset as the training images, and 20% (120 pictures) of the images as the validation images.

The performance evaluation metrics parameters for both YOLOv5s and YOLOv5s_ECA_SA_Deconv models are shown in Table 2. For all four categories of detection, the mAP@.5 of the YOLOv5s_ECA_SA_Deconv model is higher than that of the YOLOv5s model by 5.75%.

Table 2. Object detection results

Network models	Type of target	mAP@.5(%)
YOLOv5s	TOOL_A	80.5
	TOOL_B	82.6
	TOOL_C	79.5
	TOOL_D	86.1
YOLOv5s_ECA_SA_Deconv	TOOL_A	89.5
	TOOL_B	85.4
	TOOL_C	86.7
	TOOL_D	90.1

In summary, the YOLOv5s_ECA_SA_Deconv model is better than the YOLOv5s model in terms of convergence speed and detection accuracy. The effect of the YOLOv5s_ECA_SA_Deconv model in object detection in real environments is shown in Fig. 6. It can be seen that the YOLOv5s_ECA_SA_Deconv model can effectively detect the rough parts when detected in the actual dynamic environment.

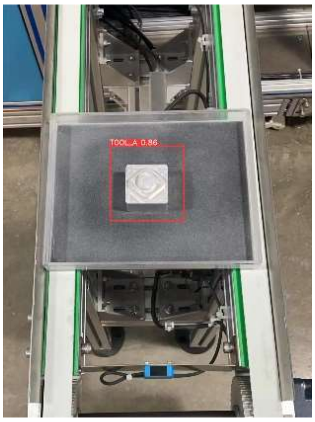


Fig. 6 Effectiveness of object detection in real-world environments

6. Conclusion

Aiming at the problems of YOLOv5s object detection algorithm when detecting in dynamic environment, an improved to carry out the improved YOLOv5s object detection algorithm is proposed, including the introduction of the hybrid attention mechanism ECA_SA module before the convolution module of the Neck network, as well as replacing the nearest-neighbor interpolation method that was originally used for up-sampling with the inverse convolution method; By comparison, the mAP@.5 of YOLOv5s_ECA_SA_Deconv is 5.75% higher than that of YOLOv5s algorithm in dynamic environment.

References

[1] JIA Shina. Research on Small Object Detection Algorithm Based on Improved YOLOv5[D]. Nanchang University, 2023 (in Chinese).

[2] ZHAO Wanyue. Study of Object Detection Algorithm Based on YOLOv5[D]. Xidian University, 2021 (in Chinese).

[3] ZHOU Yiyang. Research on mobile robot object detection algorithm based on YOLOv5[J]. Tianjin University of Technology and Education, 2021 (in Chinese).

[4] LI Xiaosong. Research on Fitness Movement Detection and Application Research Based on Improved YOLOv5s [D]. 2023 (in Chinese).

[5] Lin TY,Dollár P,Girshick R,et al.Feature Pyramid Networks for Object Detection[A].2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)[C].2017:936-944.

[6] Liu S,Qi L,Qin H,et al.Path aggregation network for instance segmentation[A].Proceedings of the IEEE conference on computer vision and pattern recognition[C].2018:8759-8768.

[7] Zheng Z,Wang P,Liu W,et al.Distance-IoU Loss:Faster and Better Learning forBounding Box Regression[J]. CoRR,2020,34(07):12993-13000.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[9] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.