Research on Low Significance Text Detection Algorithm in Text Translation

Jianyu Li
Artificial Intelligence Translation Shaanxi University Engineering Research Center, Xi'an Fanyi University, Xi'an, Shaanxi, China

Abstract: In order to address the issue of inconspicuous feature response resulting from the resemblance of certain text and background colours during the text translation process, a low saliency text detection algorithm has been proposed. Firstly, a channel spatial attention module is incorporated into the network, which concurrently assigns attention to the text target from both the channel and spatial dimensions to augment the features of the text target region. Secondly, a multi-scale feature fusion module is designed to connect different levels of information on the fusion feature map through a jump connection, thereby acquiring the fusion features with comprehensive multi-scale information. The results of the experiments demonstrate that the algorithm proposed in this paper is effective in detecting low-saliency text. Additionally, it is capable of directly outputting the enhanced image, which enhances the quality of the text translation.

Keywords: Attention Mechanism; Text Detection; Machine Translation; Multi-scale Feature Fusion

1. Introduction

The majority of contemporary text detection methodologies are founded upon object detection, which can be classified into two principal categories, the first of which is the two-stage target detection algorithm, which initially identifies the target before subsequently classifying it. The most prominent algorithms in this category are: These include R-CNN [1], Fast R-CNN [2], and R-FCN [3]. In contrast to the two-stage object detection algorithm, the single-stage algorithm is capable of locating and classifying the target simultaneously. The most notable algorithms in this category are YOLO [4], SSD [5] , and RetinaNet [6] . In comparison to two-stage object detection, one-stage detection algorithms are more expeditious, capable of meeting real-time detection requirements, and have a broader range of applications. Of these, the YOLO series, proposed by Redmon in 2016, exhibits superior performance and is the most representative. YOLO treats the detection problem as a regression problem, thereby facilitating more effective differentiation between targets and backgrounds. As the YOLO object detection algorithm has developed, in 2020, Ultralytics [7] proposed YOLOv5, which has been one of the fastest and most accurate network models among the current object detection algorithms due to the lightweight of the model [8-12].

In the process of text translation, a common and thorny problem is that some characters are similar to the background color and texture information is not obvious, which makes it difficult for the recognition algorithm to accurately capture the shape and contour of the text, which greatly affects the accuracy of text translation, and may also lead to problems such as omission, misidentification or garbled characters in the translation results. This is especially common in scanned documents, photos, or image text such as screenshots. In order to solve this problem, this paper designs a low-saliency text detection algorithm to study how to improve the accuracy of

low-saliency text in text translation. By adding a channel spatial attention module (CSAM) and a multi-scale feature fusion module (FFM) to the backbone network, the algorithm fuses different scale information processed by CSAM to obtain feature expressions with richer information and more robust scale differences, and improve the model's learning ability of the overall text features. Experiments show that the proposed algorithm greatly improves the accuracy in the detection of low-salient characters, and then reduces the error rate of text translation.

2. Low Saliency Text Detection Algorithm

The low-saliency text detection algorithm proposed in this paper is based on the YOLOv5 network as the basic framework, and the structure is shown in Figure 1, which mainly includes four parts.

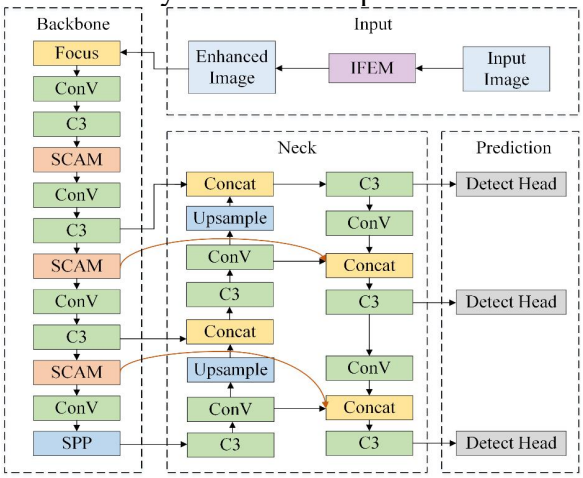


Figure 1. Network Structure of Low-Saliency Signature Detection Algorithm

In the Input part, the image feature enhancement module is designed to guide the network to predict the reasonable gamma value, and then the enhanced image is obtained through the gamma mapping function. The backbone part contains Focus, Conv, C3, SPP [13], and SCAM structures for extracting text features. Conv extracts features from the image, SCAM improves the network feature extraction ability, and provides high-quality feature maps rich in high-level information and shallow position information for the multi-scale receptive field feature pyramid. In the Neck part, FFM is designed to integrate the text feature information processed by SCAM into the feature fusion network, and realize the fusion of different scale features while learning the feature distribution between the feature map channels and spatial locations. In the prediction phase, three detection modules are employed to identify targets of varying sizes. The loss function utilises the EIOU Loss function, which is described in reference [14]. The target box is then filtered and generated through non-maximum suppression.

2.1 Channel Space Attention Module

The extraction of more accurate text feature information from the feature map represents a crucial step in the improvement of low-salient text detection performance. In this section, the channel space attention module is incorporated into the model. The input text features are compressed and weighted and space dimensions with reference to the concept of CBAM [15], thereby enhancing the network's ability to focus on bone pick text features. Figure 2 illustrates the structure of CSAM. The input feature map generates the channel weights through the channel attention mechanism, and multiplies them with the input feature map to extract more distinguishing text features, and then obtains the spatial weights through the spatial attention mechanism to capture the key details of the text target.The joint effect of the channel attention mechanism and the spatial attention mechanism enables the module to direct the network's attention towards the most salient feature channels and spatial positions within the image, thereby enhancing the network's ability to recognise low-saliency text targets with greater accuracy and resilience.

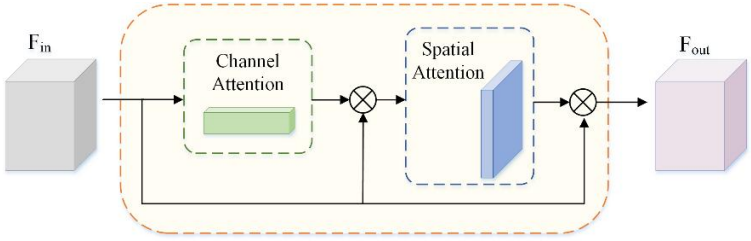


Figure 2. Structure of the Channel Spatial Attention Module

In contrast to CBAM, the channel attention output feature map is not multiplied by the spatial weight in this section. This is because the feature map Zc, which is output by the channel attention module, contains more semantic information than the input feature map. This results in a more concentrated attention area and a smaller activation value for the text target. Additionally, there is an increased probability that a low-saliency text target at the edge of the cluster region will be recognized as the background.

As illustrated in Figure 3, weights are generated from 2 simultaneous average pooling and maximum pooling. The weight assigned to each channel serves to quantify its relative importance in the context of text detection. A higher weight value indicates a greater contribution from that channel, necessitating its retention. This is calculated using the Equation (1).

$$Z_c = \sigma\left(MLP\left(AvgPool\left(Y\right)\right) + MLP\left(MaxPool\left(Y\right)\right)\right) \quad (1)$$

where σ denotes the sigmoid activation function.

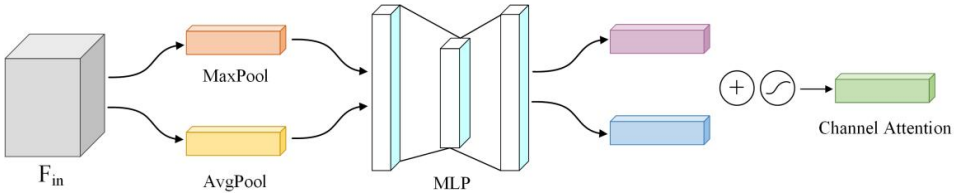


Figure 3. Channel Attention Structure

The spatial structure is illustrated in Figure 4, which predominantly employs maximum pooling and average pooling on the channel. The weighted feature map initially passes through a convolutional layer to reduce its dimension to 1/16, and then is activated through the Sigmoid function. The id function is employed to generate a spatial weight between 0 and 1. This spatial weight serves to quantify the relative importance of different pixel positions, with higher values indicating greater contributions to text detection and, consequently, greater preservation requirements. The spatial weights are then multiplied by the results of the final spatial attention map, which is calculated in accordance with the Equation (2).

$$Z_s = \sigma\left(f^{7\times7}\left(\left[AvgPool\left(Y\right); MaxPool\left(Y\right)\right]\right)\right) \quad (2)$$

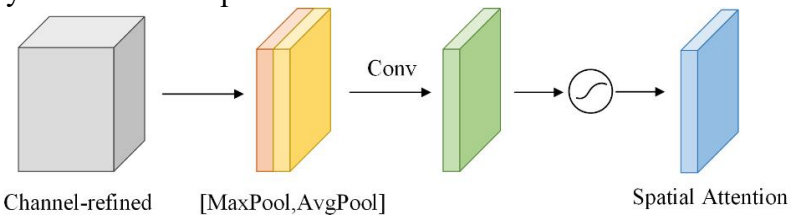


Figure 4. Structure of Spatial Attention

By learning the interrelationships between channel and spatial features, CSAM can maximize the limited feature information in bone-signature text images to help the backbone network better capture the features of low-significant bone-signature text, and make the network capable of detecting critical text information.

2.2 Multi-Scale Feature Fusion Module

As the number of layers in an object detection network increases, the network may lose some important feature information due to the layer-by-layer

abstraction inherent to the process of information transmission. Therefore, it is essential to fuse detailed feature information with abstract semantic information. Multi-scale feature fusion enables the network to conduct a more comprehensive analysis of the entire image, which is a widely used technique in object detection tasks to enhance the detection performance of network models. The common feature fusion methods are FPN [16], PANet [17], and BiFPN [18]. Among them, FPN adopts a top-down approach to fuse features at different levels to generate multi-scale feature maps, but it is limited by one-way feature information transmission. On the basis of FPN, PANet further introduces adaptive feature pooling and multi-branch fusion mechanism. The BiFPN module is a two-way weighted feature fusion module that employs a bottom-up and top-down multi-level feature fusion mechanism. It introduces a two-way path to enhance the ability of information transmission by unifying the resolution. The addition of hopping connections to the feature map, serves to mitigate the loss of feature information that can result from an excess of network layers. This modification renders the network more conducive to object detection tasks within complex scenarios. The feature fusion module can enhance the accuracy and stability of the network model by effectively integrating the feature maps.

To improve the representation of text regions with low prominence, the FFM is inspired by the BiFPN framework, as illustrated in Figure 5. Initially, the MMF unifies the dimensions of the feature map extracted from the channel spatial attention layer, completing it through a horizontal connection and downsampling operation. Subsequently, the feature information of varying levels in the channel dimension is spliced together. The incorporation of cross-scale connections enables a second stage of feature fusion, merging detail and semantic feature maps in the channel dimension. This process enriches the feature representation, ultimately producing the multi-scale fused feature map.

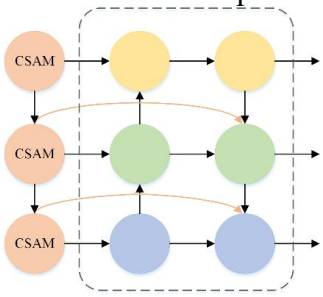


Figure 5. Structure of FFM

This paper presents FFM designed to enhance the integration of shallow detail features and deep semantic information in low-salient images. Each layer of feature information is processed by CSAM, enabling the model to learn the distribution of features across both channels and spatial locations, resulting in richer and more robust text representations. This module effectively combines global and local information from both deep and shallow levels, thereby improving the model's ability to learn overall features and enhancing its detection performance and robustness.

3. Experimental Results and Analysis

3.1 Experimental Environment and Parameter Configuration

This study was conducted on a Windows 10 64-bit system using Python 3.7, PyTorch 1.8.0, and CUDA 11.1. All models were executed on an NVIDIA RTX 3090Ti GPU, utilizing consistent hyperparameters for training, validation, and testing. The images were formatted as 640×640 JPGs, with a batch size of 8 and training spanning 200 epochs. Evaluation metrics included Precision (P), Recall (R), mean Average Precision at 0.5 (mAP0.5), weight, and actual detection speed (FPS). The formulas for calculating P and R are provided in Equations (3) and (4).

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

The term TP (True Positive) denotes the count of words that have been correctly

identified as such, while FP (False Positive) indicates the number of non-textual samples mistakenly predicted as words. FN (False Negative) represents the count of text samples incorrectly classified as non-textual. The mAP is the average class accuracy across learning, with mAP0.5 representing the average precision of the text dataset at an intersection-over-union (IoU) threshold of 0.5. The mean precision AP is derived from the area under the precision-recall curve. IThe formula for calculating mAP can be found in Equation (5).

$$mAP = AP = \int_0^1 PdR \tag{5}$$

3.2 Ablation Experiments

To assess the effectiveness of the modules, this paper performs ablation experiments on the CSAM and FFM. The influence of each module on the results is analyzed, and they are embedded into other object detection models to evaluate their generality. The findings of the ablation experiments are presented in Table 1, using the YOLOv5 algorithm as the baseline model.

Table 1. Experimental Study of Ablation

Model	P/%	R/%	mAP0.5/%	FPS/(f-S)$^{-1}$
YOLOv5	86.02	83.06	84.27	69.6
YOLOv5_CBAM	88.83	84.53	85.32	67.6
YOLOv5_FFM	88.13	86.04	86.68	65.3
YOLOv5_CBAM_FFM	90.87	87.82	89.16	58.4

3.2.1 Channel Space Attention Module

Compared to the baseline, the detection model incorporating CSAM shows a 1.05% improvement. The CBAM effectively activates key locations and semantic information in the feature map by leveraging both channel and spatial dimensions, thereby enhancing detection accuracy for low-salient text. The enhanced image emphasizes text information, keeping feature differences with the original image minimal, allowing the network to focus on this vital data while reducing noise interference. To evaluate the effectiveness of the channel spatial attention module, several popular attention mechanisms were integrated into the YOLOv5 backbone and compared with the SCAM introduced in this paper. The results are detailed in Table 2.

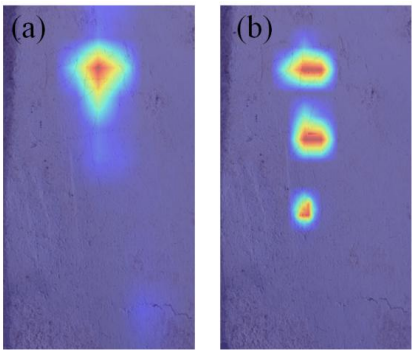
In comparison to the baseline, the detection model with CSAM has demonstrated an increase of 1.05%. The CBAM effectively activates the key location and semantic information of the target in the feature map, utilising the two dimensions of channel and space. This effectively enhances the detection accuracy of low-salient text. The enhanced image serves to highlight the text information, and the difference between its features and those of the text in the original image is minimal. This enables the network to prioritise this crucial information, while simultaneously minimising the interference of noise. In order to verify the effectiveness of the channel spatial attention module, several popular attention mechanisms were added to the YOLOv5 backbone network and compared with the SCAM added in this paper. The results are presented in Table 2.

Table 2. Performance Comparison of Adding Different Attention Mechanisms

Model	mAP0.5/%	Weight/M
YOLOv5_SE	84.39	4.1
YOLOv5_ECA	84.32	4.0
YOLOv5_CA	84.79	4.3
YOLOv5_ S^2 -MLPv2	84.74	4.3
YOLOv5_SCAM	85.32	4.1

The experimental results demonstrate that the channel spatial attention module, as incorporated in this paper, exhibits the most effective detection capabilities. After analysis, SE[19] with the ECA[20] only focuses on the weight distribution of channels, CA[21] embeds the position information into the channel information, S2-MLPv2[22] adopts a direct spatial shift operation for the feature map. SCAM considers the significance of distinct feature channels and pixels at varying positions within the same channel, concurrently

directing attention to the two dimensions.To further verify the effectiveness of CSAM, the CAM[23] method was employed to generate a heat map of the image in its original and enhanced states, as illustrated in Figure 6.



(a) no CSAM; (b) CSAM is added

Figure 6. Thermodynamic Diagram Comparison Results

Because the visual characteristics of the text are poor and too close to the background texture information, the activation value of the text target area without the CSAM is low and the activation range is too large, and the text target with no obvious features is not activated, which is difficult to effectively reflect the real text target location. After passing through the CSAM, the relatively more accurate focus on the text target area produces a higher activation value, and the activation of the background area is suppressed. It can be seen that the CSAM proposed in this paper can effectively enhance the character features of the text, and can improve the problem of detecting the target features of Chinese characters is not obvious.

3.2.2 Multi-scale feature fusion module

The mAP of the detection model improved by 2.41% after enhancing the FFM compared to the baseline. To further analyze its effectiveness, a visual comparison experiment of the feature map was conducted. Figure 7 is the result of the visual comparison experiment of the feature map of three levels in the network, which are YOLO Head1, YOLO Head2 and YOLO Head3, respectively, and the corresponding sizes are (80,80), (40,40), and (20,20). The multi-scale feature fusion module designed in this paper fuses the detailed information such as texture and edge contour extracted from the shallow layer with the deep semantic information, and produces a higher activation value in the central region of the text target, and the distribution range of the activation value is more consistent with the distribution of Chinese characters in the actual image.

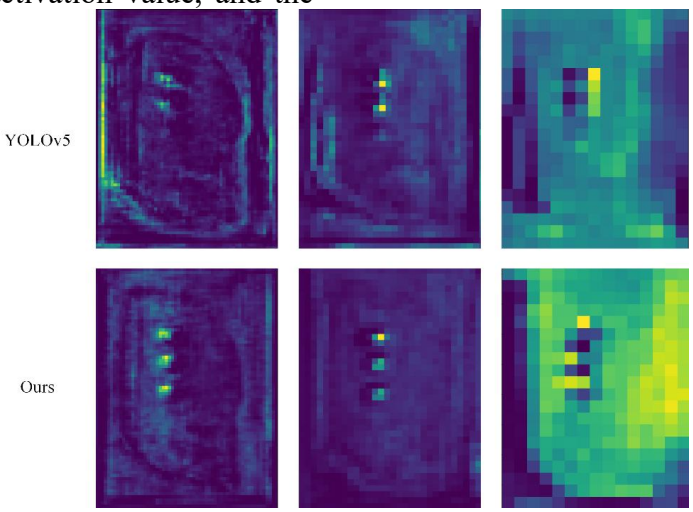


Figure 7. Visualization of Feature Map

3.2.3 Universality test of each module

Table 3 illustrates the outcomes of integrating the enhanced modules into four additional object detection models. The data in the table demonstrate that the two improved modules presented in this paper exhibit varying degrees of accuracy enhancement across the majority of object detection models. This evidence substantiates the assertion that the improved

modules in this paper possess a degree of generality.

Table 3. Universal Test of Each Module

Model	mAP0.5/%			
	RetinaNet	EfficientDet	YOLOv3	YOLOv4
_SCAM	75.68	73.53	75.12	78.02
_FFM	76.13	73.46	76.25	79.16
_CBAM_FFM	76.42	74.21	76.36	79.45

3.3 Comparative Experiments

This paper presents a comprehensive evaluation of the proposed low saliency text detection algorithm on a self-built dataset. Table 4 illustrates the accuracy of the proposed algorithm in comparison with other object detection algorithms on the same dataset and training environment. the PR curves are shown in Figure 8.

Table 4. Different Algorithms Contrast Experiment

Model	P/%	R/%	mAP0.5/%	FPS/(f-S)$^{-1}$
SSD	72.34	56.73	70.55	63.8
Faster-RCNN	63.46	87.23	80.21	30.6
RetinaNet	82.19	71.45	75.76	24.7
EfficientDet	77.53	80.03	74.02	14.8
YOLOv3[24]	83.04	67.25	74.12	61.2
YOLOv4[25]	86.24	61.55	78.15	42.7
YOLOv5	86.02	83.06	84.27	69.6
Literature [11]	84.21	79.26	83.31	60.2
Ours	90.87	87.82	89.16	58.4

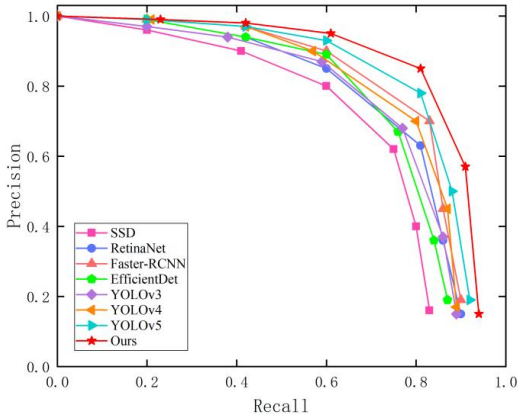


Figure 8. Contrast Experiment PR Curve

The experimental results demonstrate that the proposed algorithm achieves an mAP of 89.16% on the test set, representing a significant enhancement in detection accuracy compared to YOLOv5 and other object detection algorithms. Furthermore, the proposed algorithm exhibits superior accuracy, recall, and actual detection speed compared to other comparison algorithms under identical experimental conditions.

4. Conclusion

Due to the influence of low-salient text on the translation accuracy in the process of text translation, this paper proposes a low-salient text detection algorithm based on image feature enhancement to solve the problem that the feature response is not obvious caused by the similarity of text and background color, and establishes a text dataset, based on which the training and comparison experiments are carried out. Firstly, the channel space attention module was added in the feature extraction process, and the attention was allocated from the two dimensions of channel and space at the same time, so as to enhance the features of the text target area. Secondly, a multi-scale feature fusion module is designed to fuse information through jump connections. Experimental results show that the proposed algorithm achieves 89.16% mAP0.5 for low-saliency text detection, and has better detection performance in low-salient text detection tasks, and then obtains better text translation results.

References

[1] Girshick R, Donahue J, Darrwll T, et al. rich

feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE Conference on Computer Vision & Pattern Recognition. ieee computer society, 2014: 580-587.

[2] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[3] He KM, Gkioxari G, Dollar P, et al. Mask R-CNN[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Venice: IEEE Press, 2017. 2961-2969.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[J]. IEEE, 2016: 779-788.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//Proceedings of the European Conference on Computer Vision (ECCV),2016: 21-37.

[6] Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision(ICCV), 2017: 2980-2988.

[7] Ultrarytics. YOLOv5 [EB/OL]. (2020-6-3) [2021-4-15]. https: // github. Com / Ultraytic / YOLOv5.

[8] Liu J, Zhu X, Song MM. Improved Chinese character detection for end-to-end natural scenes in Yolov2[J]. Control and Decision 2021:2483-2489.

[9] Yin H, Zhang Z, Wang YL. Chinese Text Detection based on YOLOv3 and MSER [J]. Computer Applications and Software 2013,8(10):168-172, 195.

[10] ao MT, Sun H, Tang YQ, etc. Fingerprint secondary feature detection method based on improved YOLOv5[J]. Advances in laser and optoelectronics: 1-19[2022-07-25] .

[11] Dong YS, Li ZX, Guo JY, etc. An improved YOLOV5 x-ray contraband detection model [J]. Advances in laser and optoelectronics: 1-17[2022-07-25] .

[12] Feng Z, Xie Z, Bao Z, et al. Real-time dense small target detection algorithm for uav based on improved YOLOv5[J]. Acta Aeronaut. Sin, 2022: 1-15.

[13] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence 2015, 37(9):1904-1916.

[14] Zhang YF, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J/OL]. arXiv:2101.08158v1.

[15] Woo S, Park J, Lee JY, et al. CBAM: Convolutional block attention module [C] // Proceedings of the Europe-an conference on computer vision. Munich. Germany: springer, 2018:3-19.

[16] Lin TY, Dollár P, et al. Feature pyramid networks for object detection [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017,5936-5944.

[17] [17] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018, 1701-1709.

[18] Tan MX, Pang RM, Quoc V. Le. EfficientDet: scalable and efficient object detection[J]. arXiv:1911.09070 [cs.CV].

[19] Hu J, Li S. Squeeze and excitation networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence 2017, 99(10):7132-7141.

[20] Wang QL, Wu BG, Zhu PF, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. ieee,2022:11531-11539.

[21] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design [C] //IEEE. Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. atlanta: ieee, 2021:4510-4522.

[22] Yu T, Li X, Cai YF, et al. S2-MLPv2: Improved spatial-shift mlp architecture for vision[C]//IEEE conference on Computer Vision and Pattern Recognition. 2021, arXiv:2108.01072 [cs.CV].

[23] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 2921-2929.

[24] Redmon J, Farhadi A. Yolov3: An incremental improvement[C]//IEEE conference on Computer Vision and Pattern Recognition, 2018, arXiv: 1804.0276.

[25] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: optimal speed and accuracy of object detection[C]//IEEE conference on Computer Vision and Pattern Recognition. 2020. arXiv:2004.10934v1 [cs.CV].

