MSFYOLO: Feature Fusion-Based Detection for Small Objects

Ziying Song, *Member, IEEE,* Yu Zhang, Yi Liu, Kuihe Yang and Meiling Sun

***Abstract*—At present, the effect of object detection algorithm in small object detection is very poor, mainly because the low-level network lacks semantic information and the characteristic information expressed by small object inspection data is very lack. In view of the above difficulties, this paper proposes a small object detection algorithm based on multi-scale feature fusion. By learning shallow features at the shallow level and deep features at the deep level, the proposed multi-scale feature learning scheme focuses on the fusion of concrete features and abstract features. It constructs object detector (MSFYOLO) based on multi-scale deep feature learning network and considers the relationship between a single object and local environment. Combining global information with local information, the feature pyramid is constructed by fusing different depth feature layers in the network. In addition, this paper also proposes a new feature extraction network (CourNet), through the way of feature visualization compared with the mainstream backbone network, the network can better express the small object feature information. The proposed algorithm is valuated on the MS COCO and achieved leading performance with 11.7% improvement in FPS, 17.0% improvement in AP, 81.0% improvement in ARS, and 23.3% reduction in computational FPLOs compared to YOLOv3. This study shows that the combination of global information and local information is helpful to detect the expression of small objects in different illumination. MSFYOLO uses CourNet as the backbone network, which has high efficiency and a good balance between accuracy and speed.**

***Index Terms*—Object detection, Feature extraction network, Feature pyramid, Multi-scale feature fusion**

I. Introduction

Object detection is an important research direction in computer vision field [1]. Since 2012, AlexNet [2] has won the champion in the image classification task of ILSVRC Challenge [3] with a significant advantage, which shows the strong feature representation ability of convolutional neural network. Since then, the research boom based on deep learning has begun. In the following years, more powerful classification networks such as VGGnet [4], Googlenet [5], ResNet[6], Inception v3[7], Densenet [8], Senet [9], etc., were introduced. Because they are able to extract very abstract features, they are also commonly used as skeleton networks for more complex computer vision tasks, including object detection, in addition to image classification. In 2014,R-CNN algorithm[10] surpasses DPM algorithm[11] with an absolute advantage in PSACAL VOC detection data set [12], marking a new milestone for object detection. Since then, deep learning algorithm has occupied an absolute dominant position in the research field of object detection and has continued to this day [13].

Whether selecting a one-stage object detection method or a two-stage object detection method, it is difficult to detect small object. Specifically, small object detection mainly faces the following challenges:

Firstly, the underlying features lack semantic information [14]. In the existing object detection model, the underlying features of Backbone are generally used to detect small objects, but the lack of semantic information of the underlying features brings some difficulties to the detection of small objects.

Secondly, the amount of training sample data for small objects is small [15]. Since the collected sample pictures need to be manually marked, which makes it impossible to fully learn the small and medium-sized objects in the process of model training.

Thirdly, the difference between the Backbone used in the test model and the test task. Backbones of the existing object detection model are all trained on the classification data set, but the scale distribution of the object in the classification data set is different from that in the detection dataset [16].

II. Materials and Methods

A. *CourNet Extracting Feature Networks*

Structure of YOLOv5 Backbone was modified into a C3 network structure (Fig. 1). The reduction of the convolutional layer extracted shallow layer charateristic information better than the BottlenneckCSP structure.

Layers of C3 network structure and two layers of Conv network structure are added alternately.

The same point between the backbone of YOLOv5l and CourNet is that the last layer processes features through SPP [17] (Fig. 2).

In the general CNN structure, full connection is usually connected behind the convolution layer. The number of features in the full connection layer is fixed, so the input size will be

Z. Song (*Member, IEEE*) was with Hebei Normal University of Science and Technology, Qinhuangdao, Hebei, CO 066000 China, in 2019. He is now with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, CO 050018 China (e-mail: songziying@stu.hebust.edu.cn).

Y. Zhang, is now with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, CO 050018 China (e-mail: zhangyu@stu.hebust.edu.cn).

Y. Liu, is now with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, CO 050018 China (e-mail: liuyi@stu.hebust.edu.cn).

K. Yang, is now with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, CO 050018 China (e-mail: ykh@hebust.edu.cn).

M. Sun, is now with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, CO 050018 China (e-mail: sunmeiling@stu.hebust.edu.cn).

fixed when the network is input. The size of our input image is always unable to meet the required size. However, the usual techniques are crop and warp. However, the ratio aspect of the image and the size of the input image are changed. This distorts the original image. SPP layer can solve this problem well, usually connected to the last layer of the convolution layer.

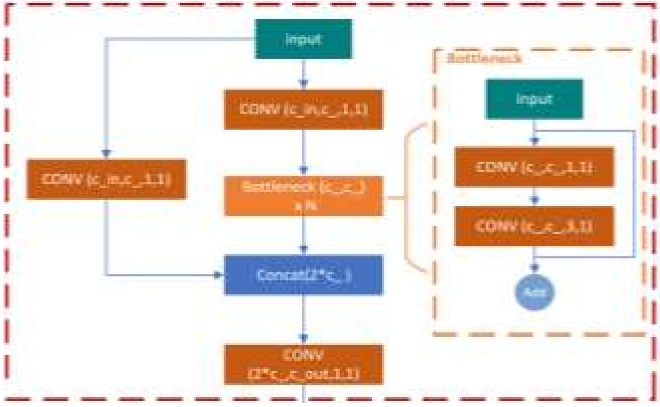


Fig. 1. C3 network structure.

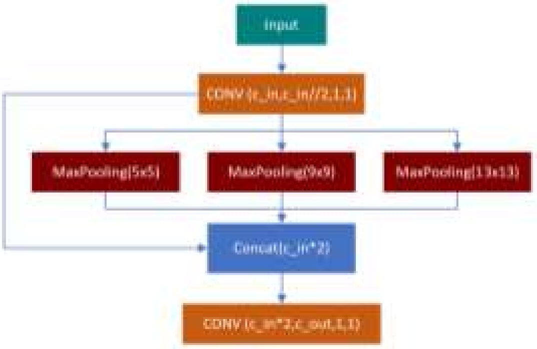


Fig. 2. SPP network structure.

When using CNN to detect objects, the shallow network has small receptive field and high resolution, and the low features are more suitable for small object or simple object detection because of their rich details. Deep network has large receptive field and low resolution, and its high-order features are more suitable for large or complex object detection[36] because of their rich semantic information. Because the shallow network can better represent the semantic feature information, and the information of small object is semantic information. Therefore, if the feature extraction network can obtain better semantic information and edge detection information in the shallow network, the feature extraction network has better extraction effect on semantic information, and is more suitable as the backbone of small object detection. Through the comparison of feature visualization: VGG19[18], ResNet152[19], Inception v3[20] and CourNet, the comparison diagram of feature visualization is shown in Fig. 3. Among them, Fig. 3 (a) is the original input image, Fig.3 (b) is the feature image generated by VGG19, Fig. 3 (c) is the feature image generated by Resnet152, Fig. 3 (d) is the feature image generated by YOLOv5 backbone, Fig.3 (e) is the feature image generated by Inception v3, and Fig. 3 (f) is the feature image generated by CourNet. It can be seen from the Figure that the feature graphs of Fig. 3 (c), Fig. 3 (d) and Fig. 3 (f) perform better in semanticsYOLOv5 Backbone.

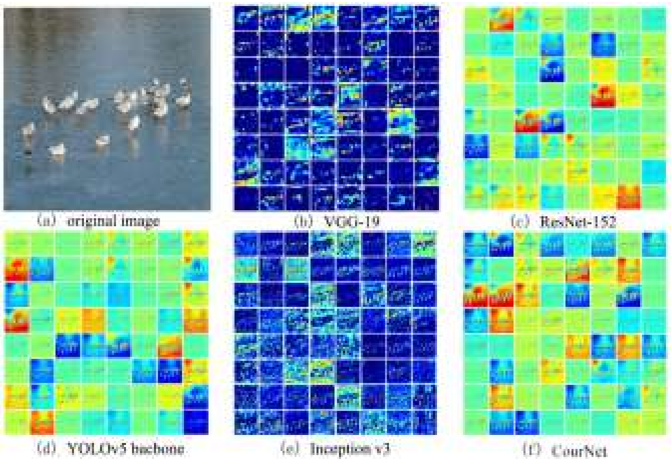

(a) original image
(b) VGG-19
(c) ResNet-152
(d) YOLOv5 backbone
(e) Inception v3
(f) CourNet

Fig. 3. Comparison chart of feature visualization.

B. Multi-Scale Feature Fusion

MSFYOLO adopts the path aggregation network (PANET) structure [21] as the model architecture of neck, as shown in Fig. 4. PANET starts from the bottom of the feature pyramid that FPN [22] has built, and adds a side path of feature re-fusion from bottom to top, which reconstructs a pyramid that strengthens spatial information. Then, the ROI alignment operation is applied to each layer of the pyramid, and the aligned feature layers are fused by taking the maximum value, and finally detected on the fused feature graph, to ensure that the prediction of each object makes full use of the information of all feature layers. The reason why PANET is selected in MSFYOLO is that it can accurately save spatial information, help to correctly locate pixels and form a mask.

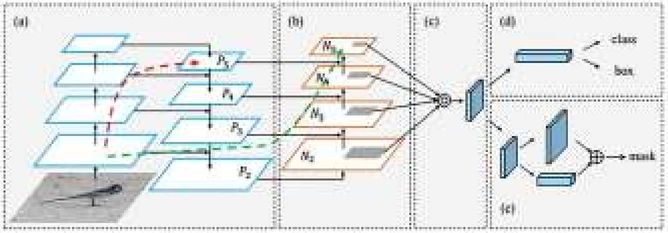


Fig. 4. PANET structure.

In the detection of 640×640 input image, the mesh sizes of the three feature layers of YOLOv5 are 80×80, 40×40 and 20×20, respectively [23]. The deeper the convolutional neural network is, the larger the Receptive Field of the characteristic map is, which also means that each neuron contains more global and high-level semantic features, but local features and detailed features will be lost. On the contrary, when the convolutional neural network is shallow, the features contained in the neurons of the feature map tend to be more local and detailed [24]. In order to better identify small object, a detection layer is added compared with the detection layer of YOLOv5. The new detection layer is used to detect 4×4 pixel object. The sizes of the four feature layers of MSFYOLO algorithm are 160×160, 80×80, 40×40 and 20×20 respectively. The network structure of MSFYOLO algorithm is shown in Fig. 5.

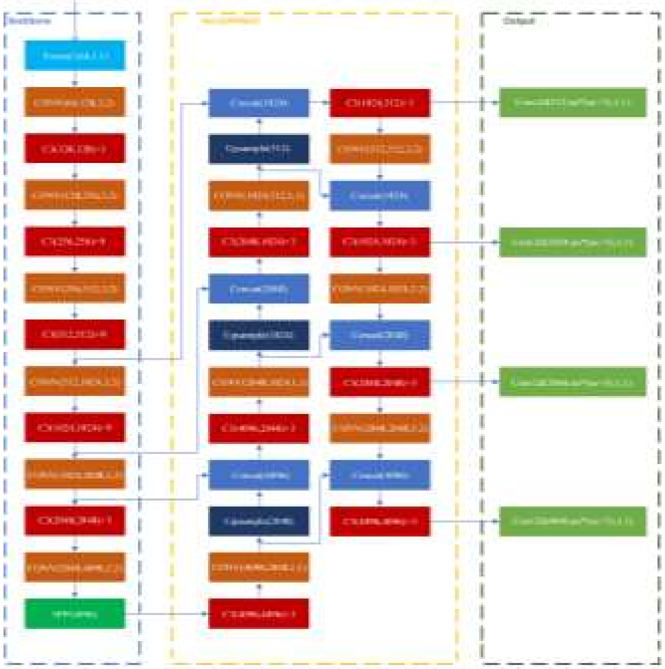


Fig. 5. The network structure of MSFYOLO algorithm.

C. MSFYOLO Loss Function

It is what we call cross and contrast ratio, which is the most used index in object detection, not only used to determine positive samples and negative samples, but also used to evaluate the distance between the output box and the real box [25]. The calculation formula is shown in equation 1, where A and B are the areas of two object boxes respectively. It can reflect the detection effect of the predictive test box and the real test box. Another good feature is scale invariance.[35] In the regression task, the most direct indicator to judge the distance between the output box and the real test box is IoU [26]. If two boxes do not intersect = 0, cannot reflect the distance between them. Meanwhile, since loss = 0, there is no gradient return, so learning and training cannot be carried out. The loss function of is shown in equation 2.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$\mathcal{L}_{IoU} = -ln(IoU) \tag{2}$$

Since IoU is the concept of ratio, it is insensitive to the scale of the object object. However, the BBox regression loss optimization and IoU optimization in the detection task are not completely equivalent, and the Ln norm is also sensitive to the scale of the object, so IoU cannot directly optimize the non-overlapping part [27]. GIoU is the lower bound of IoU, in the case of infinite overlap between the two boxes, GIoU =1. IoU value [0,1], but GIoU has a symmetric interval, value range [-1,1]. The maximum value is 1 when the two overlap, and the minimum value is -1 when they have no intersection and are infinitely far away. Therefore, it is a very good distance metric [28]. $GIoU$, is also a measure of distance, as shown in equation 3. Where, is the minimum closure area of two object boxes, and is the intersection area of two object boxes. $A^c U$ The loss function of GIoU is shown in equation 4.

$$GIoU = IoU - \frac{A^c - U}{A^c} \tag{3}$$

$$\mathcal{L}_{GIoU} = 1 - GIoU \tag{4}$$

DIoU is more consistent with the mechanism of object box regression than GIoU, taking into account the distance, overlap rate and scale between the object and anchor, making object box regression more stable and avoiding divergence and other problems in the training process like IoU and GIoU. Similar GIoU Loss, DIoU Loss can still provide the direction of movement for the bounding box when it does not overlap with the object box. DIoU Loss directly minimizes the distance between the two object boxes and therefore converges much faster than GIoU. For the case containing two boxes in the horizontal and vertical directions, the DIoU loss can make the regression very fast, while the GIoU loss almost degenerates to the IoU loss [29]. DIoU can replace IoU evaluation strategy and be applied to NMS[30] to make the results obtained by NMS more reasonable and effective. DIoU calculation is shown in equation 5. The loss function of DIoU is shown in equation 6.

$$DIoU = IoU - \frac{\rho^2(b, b^{gt})}{c^2} \tag{5}$$

$$\mathcal{L}_{DIoU} = 1 - DIoU \tag{6}$$

The length-width ratio of the three elements of Bbox regression has not been taken in the calculation, so CIoU is further proposed on the basis of DIoU[31]. The calculation formula is shown in equation 7[32]. The loss function of CIoU is shown in equation 8.

$$CIoU = DIoU + \alpha\vartheta \tag{7}$$

$$\mathcal{L}_{CIoU} = 1 - \text{CIoU} \tag{8}$$

Formula symbols used in the table shown in Table 1.

TABLE I
SPECIFIC PARAMETERS OF MSFYOLO

Symbol	Notes
IoU	The most commonly used index in object detection is not only used to determine positive samples and negative samples, but also to evaluate the distance between the output box and the real box.
A、B	A、B is the area of two object boxes respectively.
$\mathcal{L}_{IoU}$	$\mathcal{L}_{IoU}$ *is* IoU using the log loss function.
ρ	It represents European distance.
b	It represents the center point of the object box of the predicted value.
b^{gt}	It represents the center point of the object box of the real value.
c	It the diagonal length of the minimum intersection box between the real value and the predicted value of the object box
$\alpha\vartheta$	$\alpha\vartheta$ is the weight function, and the similarity used to measure the aspect ratio is defined as $\alpha\vartheta = \frac{4}{\pi^2}(arctan\frac{\omega^{gt}}{h^{gt}} - arctan\frac{\omega}{h})$

The differences of each loss function were compared comprehensively: the overlapping area of detection box and object box was mainly considered. On the basis of IoU, solve the problem when the bounding box does not coincide [33]. Based on IoU and GIoU, the information of the center point distance of the bounding box is considered.Based on the DIoU, the scale information of the aspect ratio of the boundary box is considered [34].

The regression method adopted in YOLOv5 makes the prediction box regression faster and more accurate [35]. In this way, three important geometric factors should be taken in the object box regression function: overlapping area, distance from center point, and aspect ratio [36].

III. Experiments, Results, and Analysis

This experiment is based on the following hardware configuration: Intel i7-9750h CPU, 256g memory, 32GB video memory, NVIDIA Tesla V100. The corresponding software configuration includes windows 10 system, CUDA10.01 and python 3.7. The detection effect of MSFYOLO is greatly influenced by the detection parameters, which are shown in Table 2.

TABLE II
Specific Parameters of MSFYOLO

Name	Value	Notes
Train batch	16	When training the model, 16 images are trained in 1 time, and the weights are updated once per batch of samples.
Train subdivisions	4	If the computer memory is not large enough, divide the batch into subdivisions, each sub-batch size is batch/subdivisions
width	608	Height of input image
height	608	The width of the Input image
channels	3	Number of channels of Input image
momentum	0.949	Hyperparameters of the momentum gradient descent method, in order to solve the Hessian matrix pathological condition problem
decay	0.0005	Weight decay regular term to prevent overfitting
saturation	1.5	Generate more training samples by adjusting the saturation
exposure	1.5	Generate more training samples by adjusting exposure
hue	0.1	Generate more training samples by adjusting hue
learning_rate	0.001	Initial learning rate
burn_in	1000	If the batch is larger than burn_in, the learning rate is applied in a policy way.
max_batches	20000	Stop learning after training reaches max_batches
policy	steps	Policy to adjust the learning rate
cutmix	1	Data enhancement policy, 1 means use, 0 means don't use
activation	mish	The activation function is mish
num	9	Each grid cell predicts several boxes, consistent with the number of anchors. When you want to use more anchors, you need to increase the num, and if the training Obj tends to 0 after increasing the num, you can try to increase the object_scale.
jitter	0.3	Suppress overfitting by adding noise through dithering
ignore_thresh	0.7	This parameter determines whether the IOU error needs to be calculated to be greater than the threshold, and the IOU error is not restricted to the cost function.
truth_thresh	1	When ignore_thresh is too large, it approaches 1 and then engages. The number of regressions in the detection box will be less and can easily lead to overfitting.
iou_thresh	0.213	If ignore_thresh is set too small, the number of participants involved in the calculation will be large. Also, it is easy to cause underfitting when performing the detection frame regression.
a_t	0.25	When model prediction is performed, the
r	2	use batch normalization to prevent overfitting. Where: 1 means use; 0 means do not use

This paper intends to use MS COCO 2017 (Microsoft COCO: common objects in context 2017) data set for training and evaluation. The data set is not used with other public image data sets, and there are more small-scale images, which is more suitable for the training and testing of small object detection algorithm[37]. The data set contains 123287 images, which are divided into training set and verification set. In order to facilitate our experiments, some image enhancement methods are used to expand the size of the data set to 700000 images, including random cropping, brightness / contrast / saturation enhancement[38]. In addition, we use cutmix image enhancement strategy. The data set covers a wide range of categories (Fig. 6), in line with our research requirements. The evaluation standard of this data set is very strict[39]. Compared with other data sets, the average accuracy of common object detection algorithms using this data set will be lower [40].

Fig. 6. Example image of MS COCO.

A. mAP, mAR, FPS and FLOPs on MS COCO

Comparing SSD300, YOLOv4, Faster RCNN, YOLOv3, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x on MS COCO dataset, this training is based on trainval35k, and we have tested about 20K images on the test development set. The test results are shown in Table 3 and Table 4. In this leading board, MSFYOLO models obtain leading performance in terms of AP and achieve top AP in the object class of train. MSFYOLO algorithm performs best in AP, AP50, AP75, APS, APM, AR1, AR10, ARS and arm, especially in APS and APS. Because there are more small objects in MS COCO dataset, MSFYOLO is more inclined to detect small objects. We can see that MSFYOLO does not perform well in the detection of large objects, but it performs better in the detection of small objects. When we do multi-scale feature fusion, we prefer to express the specific features of small objects.

TABLE III
AP on MS COCO test-dev 2017

Algorithm	AP	AP50	AP75	APS	APM	APL
RetinaNet	31.4	52.4	34.2	22.4	40.3	49.1
SSD300	26.6	43.7	27.2	9.4	25.9	41.4
Faster RCNN	28.7	47.6	28.5	20.3	39.5	45.3
YOLOv3	27.8	46.3	29.1	16.8	34.7	42.9
YOLOv4	30.1	49.3	30.4	19.5	38.7	42.3
YOLOv5s	30.4	51.3	31.1	19.9	41.6	46.5
YOLOv5m	30.7	51.4	31.9	21.3	42.7	46.1
YOLOv5l	32.1	52.6	33.4	21.8	43.8	48.3
YOLOv5x	31.6	52.1	32.7	22.3	43.3	48.6
MSFYOLO	33.5	52.9	34.9	23.6	44.1	48.5

TABLE IV
AR ON MS COCO TEST-DEV 2017

Algorithm	AR1	AR10	AR100	ARS	ARM	ARL
RetinaNet	26.2	45.9	50.3	24.4	51.2	64.5
SSD300	23.7	35.1	37.2	11.6	40.4	59.4
Faster RCNN	24.2	47.8	42.8	20.5	45.2	60.2
YOLOv3	26.4	39.8	39.5	15.3	43.8	62.3
YOLOv4	25.1	40.4	42.9	18.7	47.5	64.4
YOLOv5s	26.4	42.3	44.8	20.3	49.3	64.9
YOLOv5m	26.7	43.3	48.7	23.4	48.4	65.3
YOLOv5l	27.6	44.9	47.7	26.7	48.1	63.6
YOLOv5x	26.9	44.6	49.3	25.3	49.2	62.3
MSFYOLO	27.8	44.7	49.6	27.7	49.7	64.7

The performance test results of FPS and FLOPs are shown in Table 5. We use the same experimental environment to test all the algorithms. GPU uses a NVIDIA Tesla V100 32GB. SSD300 as the largest FPS, followed by MSFYOLO (320 × 320). MSFYOLO grasps the speed and detection performance. Under the same resolution (640×640), the FPS of MSFYOLO (Focus + CourNet) is 55.3, which is the highest. MSFYOLO compares different backbones, among which Focus+CourNet is the highest in FPS. With the increase of resolution, FPS becomes smaller. When FPS is greater than 30, it is a prerequisite for automatic driving. When FPS is greater than 60, it is a game level requirement. At present, there are only two models which FPS is greater than 60, respectively SSD300 and MSFYOLO (Focus+CourNet), but the performance of SSD300 is almost the penultimate. The FPS of MSFYOLO (1280× 1280) is 23.1, which is the penultimate. Among all the algorithms, the highest FLOPs is MSFYOLO (1280 × 1280), followed by Faster RCNN. Algorithm complexity includes spatial complexity and temporal complexity. The time complexity, called the number of operations of the model, is measured by FPLOPs, and the space complexity, called the number of visits, is measured by the number of parameters.

TABLE V
FPS AND FLOPS ON MS COCO TEST-DEV 2017

Algorithm	Resolution	backbone	FPS	FLOPs	Space/ MB
RetinaNet	640× 640	ResNet101	49.4	14.4G	278.3
SSD300	300× 300	VGG16	71.2	31.4G	262.0
Faster RCNN	512× 512	ResNet101	9.8	73.0G	185.5
YOLOv3	416× 416	Darknet-53	49.5	36.1G	240.0
YOLOv4	608× 608	CSPDarknet53	42.9	18.7G	246.4
YOLOv5s	640× 640	CSPDarknet53	54.8	16.3G	7.3
YOLOv5m	640× 640	CSPDarknet53	49.5	23.4G	21.4
YOLOv5l	640× 640	CSPDarknet53	47.4	46.7G	47.0
YOLOv5x	640× 640	CSPDarknet53	45.9	85.3G	87.7
MSFYOLO	640× 640	CourNet	55.3	37.8G	50.9
MSFYOLO	640× 640	CSPDarknet53	49.8	41.2G	92.1
MSFYOLO	640× 640	CSPDarknet53+ SElayer	43.3	54.3G	105.8
MSFYOLO	640× 640	ResNet101	41.1	27.7G	192.6
MSFYOLO	320× 320	CourNet	61.6	29.8G	50.9
MSFYOLO	1280 × 1280	CourNet	23.1	96.9G	50.9

B. *Visualization of Test Results on MS COCO*

In order to verify the effectiveness of small object detection, we conducted comparative experiments on YOLOv3, RetinaNet, YOLOv5l and MSFYOLO, as shown in Fig. 7.

4

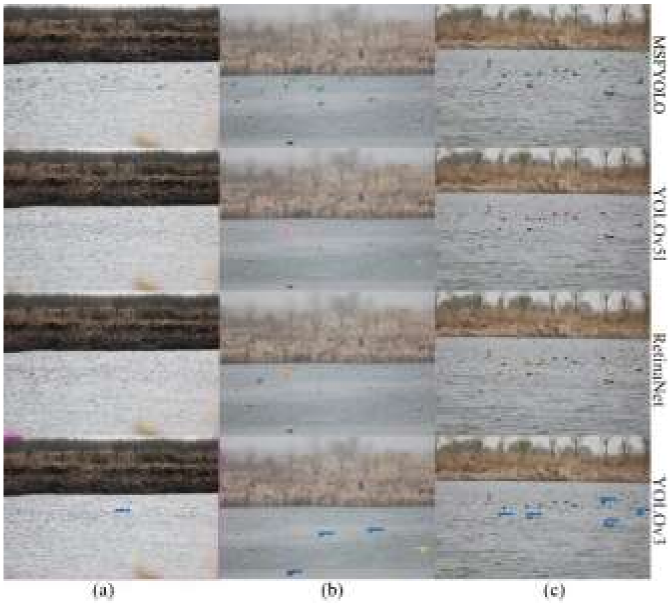


Fig. 7. Comparison of experimental results. (a) Images seriously affected by light, (b) Images affected by fog, (c) Images taken normally.

Fig. 7 (a) in the experimental comparison results, MSFYOLO most correct test result, the second is YOLOv5l and RetinaNet YOLOv3 correct test results at least, is affected by light, the MSFYOLO, YOLOv5l, RetinaNet and YOLOv3 four set of algorithms were error detection of Birds, MSFYOLO, YOLOv5l algorithm and RetinaNet three groups are the small grebe error detection to the seagulls, YOLOv3 the entire image detection for egrets. Among the undetected bird objects, MSFYOLO had the least number of undetected bird objects, followed by YOLOv5l and RetinaNet, and YOLOv3 had the most undetected bird objects. Conclusion: In the bird test data set greatly affected by light, the number of correct detection is the largest and the number of undetected is the least compared with the three algorithms of YOLOv5l, RetinaNet and YOLOv3. The number of error detection is equal to that of the three algorithms of YOLOv5l, RetinaNet and YOLOv3, and the model performs better.

Fig. 7(b) in the experimental comparison results, MSFYOLO most correct test result, the second is YOLOv5l and RetinaNet YOLOv3 correct test results at least, MSFYOLO, YOLOv5l and RetinaNet Birds object of three groups of algorithm has error detection, small MSFYOLO will grebe mistakenly identified as seagulls and small grebe, YOLOv5l and RetinaNet will be a small grebe mistakenly identified as two small grebes, YOLOv3 no error detection. Among the undetected bird objects, YOLOv3 was the highest, followed by RetinaNet and YOLOv5l, and MSFYOLO had the least detected bird objects. Conclusion: In Fig. 7(b), the number of object boxes correctly detected by MSFYOLO is the most, the number of undetected object boxes is the least, and the number of wrongly detected object boxes is more than that of YOLOv3.

The experimental comparison results in Fig. 7(c), the correct detection results of MSFYOLO are the most, followed by the correct detection results of YOLOv5l and RetinaNet, and the correct detection results of YOLOv3 are the least. Among the undetected bird objects, MSFYOLO had the best performance, followed by YOLOv5l and RetinaNet, and YOLOv3 had the most undetected bird objects. In the four algorithms of MSFYOLO, YOLOv5l, RetinaNet, and YOLOv3, two fricasees are detected by mistake as one dabchick. Conclusion: MSFYOLO algorithm performs best in the correct detection and undetected statistics, but only MSFYOLO has the case of false detection.

IV. Conclusion

In the past, more small object detection algorithms focused more energy on the way based on data enhancement, but through the combination of deep network and shallow network and multi-scale prediction, they gave more loss functions to large-scale object detection and ignored small objects. In order to realize the accurate detection of small objects, we focus more on the expression of small objects physical information through multi-scale detection, and propose a small objects detection algorithm MSFYOLO based on multi-scale feature fusion. K-means algorithm is used to cluster MS COCO samples, and a priori box parameters of different sizes are obtained. The feature extraction network CourNet is proposed, and the multi-scale feature fusion in PANET is proposed to improve the detection level and detection ability of small objects. Compared with similar algorithms with the best performance, such as YOLOv5 and RetinaNet, MSFYOLO is in a leading position in accuracy and speed, can better express the specific information of small objects, and is in a leading position in grasping the details of small objects under different lighting conditions.

The proposed algorithm MSFYOLO integrates deeper backbone networks and more scales for small objects prediction. We use the idea of feature pyramid to express the semantic features of deep network and the physical information of shallow network more reasonably. However, it also brings an urgent problem, that is, the network parameters become larger and the amount of calculation is larger due to the deepening of the backbone network.

This algorithm is similar to the current algorithm in that it uses an anchor-based model and a single-stage object detection algorithm approach. The biggest difference is that we propose a new backbone network CourNet, which is more inclined to the feature representation of semantic information and passes shallow semantic information to the tail of the network by means of PANet pyramids.

In the future, we will improve the backbone network and grasp the detection speed while ensuring the overall detection accuracy. We will also perform semantic segmentation of small objects of MS COCO on MSFYOLO.

Acknowledgment

The authors gratefully acknowledge the insightful comments and suggestions of the reviewers, which have improved the presentation. The authors sincerely thank Hebei University of Science and Technology. The authors sincerely thanks to the computer vision research laboratory of Hebei University of science and technology.

Conflicts of Interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Data Availability

The data used to support the results of this study can be obtained from the corresponding authors on request. The dataset we use is MS COCO 2017 public dataset, which can also be downloaded from https://cocodataset.org/#download.

Code Availability

The code of the paper is public, and the download address is https://github.com/modaxiansheng/Birds-YOLO.

AGRADECIMIENTOS

The submission date of this paper is August 11, 2021. This work was supported in part by the Local Cooperation Project of China Scholarship Committee under Grant 201808130283 and in part by the of Artificial Intelligence Collaborative Education Project of the Ministry of Education of China under Grant 201801003011.

References

[1] S. Wu, Y. Xu, D. Zhao, “Overview of Object Detection Based on Deep Convolutional Networks,” *Pattern Recognition and Artificial Intelligence*, 2018: pp. 335–346.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, 25, pp. 1097–1105.

[3] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei–Fei, “Imagenet: A large–scale hierarchical image database,” 2009 *IEEE conference on computer vision and pattern recognition. IEEE*, 2009: pp. 248–255.

[4] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large–scale image recognition,” arXiv preprint arXiv: pp.1409-1556, 2014.

[5] C. Szegedy, W. Liu, Y.Jia, “Going deeper with convolutions,” *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015: pp. 1–9.

[6] K. He, X. Zhang, S. Ren, “Deep residual learning for image recognition,” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: pp. 770–778.

[7] C. Szegedy, V. Vanhoucke. S. Ioffe, “Rethinking the inception architecture for computer vision,” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: pp. 2818–2826.

[8] G. Huang, Z. Liu, L. Van, “Densely connected convolutional networks,” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: pp. 4700–4708.

[9] J. Hu, L. Shen, G. Sun, “Squeeze–and–excitation networks,” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: pp. 7132–7141.

[10] R. Girshick, J. Donahue, T. Darrell, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: pp. 580–587.

[11] M. Everingham, L. Van, C. K. I. Williams, “The pascal visual object classes challenge,” *International journal of computer vision*, 2010, 88(2): pp.303–338.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, “Object detection with discriminatively trained part–based models*,” IEEE transactions on pattern analysis and machine intelligence*, 2009, 32(9): pp.1627–1645.

[13] S. Wu, Y. Xu, D. Zhao, "Overview of object detection based on deep convolutional network," *Pattern Recognition and Artificial Intelligence*, 2018 (04, 2018): pp. 335–346.

[14] Z. Zhang, X. Zhang, C. Peng, "Exfuse: Enhancing feature fusion for semantic segmentation," *Proceedings of the European Conference on Computer Vision*. 2018: pp. 269–284.

[15] X. Pan, X. Zhang, W. Dong, H. Yao, C. Xu, "Research Status of Small Sample Object Detection," *Journal of Nanjing University of Information Science & Technology*,2019,11(06): pp. 698–705.

[16] J. Yuan, Y. Hu, Y. Sun, "Deep learning for small object detection," *Journal of Beijing University of Technology*, 2021, 47(3) pp. 293–302.

[17] J. Wang, J. Zhang, J. Zhang, "A new method for image classification and object detection using convolutional neural networks," *Computer Engineering and Applications*, 2017, 53(13) pp. 34–41.

[18] A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[19] K. He, X. Zhang, S. Ren, J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), 37(9): pp.1904–1916,2015.

[20] S. Liu, L. Qi, H. Qin, "Path aggregation network for instance segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: pp. 8759–8768.

[21] Z. Zheng, P. Wang, W. Liu, "Distance–IoU loss: Faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, 34(07): pp. 12993–13000.

[22] T. Y. Lin, P. Dollár, R. Girshick, "Feature pyramid networks for object detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: pp. 2117–2125.

[23] H. Rezatofighi, N. Tsoi, J. Y. Gwak, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: pp. 658–666.

[24] S. H. Chen, C. C. Tsai, "SMD LED chips defect detection using a YOLOv3-dense model." Advanced Engineering Informatics, 2021,47 (5): pp. 101255–101259.

[25] B. Krawczyk, (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), pp. 221-232.

[26] H. Lee, M. Park, & J. Kim. (2016, September). Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In 2016 *IEEE International Conference on Image Processing* (ICIP) (pp. 3713-3717). *IEEE.*

[27] J. M. Johnson, & T. M. Khoshgoftaar, (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27.

[28] R. A. Bauder, & T. M. Khoshgoftaar, (2018). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. Health Information Science and Vystems, 6(1), p 9.

[29] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy,"Training deep neural networks on imbalanced data sets," *in Proceedings of the International Joint Conference on Neural Networks*, 2016, doi: 10.1109/IJCNN.2016.7727770.

[30] D. Verma, C. Bose, N. Tufchi, K. Pant, V. Tripathi, and A. Thapliyal, "An efficient framework for identification of Tuberculosis and Pneumonia in chest X-ray images using Neural Network," *in Procedia Computer Science*, 2020, doi:10.1016/j.procs.2020.04.023.

[31] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection ofCOVID-19 in X-Rays using nCOVnet," Chaos, Solitons and Fractals, 2020, doi: 10.1016/j.chaos.2020.109944.

[32] Y. Liang, G. Wang, W. Li, "A New Object Detection Method for Object Deviating from Center or Multi Object Crowding." Displays, 2021,69(9): 102042–102049.

[33] E. Cortés and S. Sánchez, "Deep Learning Transfer with AlexNet for chest X-ray COVID-19 recognition," *in IEEE Latin America Transactions*, vol. 19, no. 6, pp. 944-951, June 2021, doi: 10.1109/TLA.2021.9451239.

[34] O. L. V. de Sousa, D. M. V. Magalhães, P. de A. Vieira and R. Silva, "Deep Learning in Image Analysis for COVID-19 Diagnosis: a Survey," *in IEEE Latin America Transactions*, vol. 19, no. 6, pp. 925-936, June 2021, doi: 10.1109/TLA.2021.9451237.

[35] M.Rostami, K.Berahmand, and S.Forouzandeh. "A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. " J Big Data 7 ,October 2020,doi: 10.1186/s40537-020-00352-3.

[36] M.Rostami, K.Berahmand, and S.Forouzandeh. "A novel community detection based genetic algorithm for feature selection." J Big Data 8, January 2021.,doi:10.1186/s40537-020-00398-3.

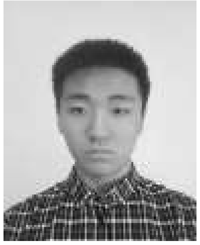
[37] M.Rostami, K.Berahmand, N.Elahe and S.Forouzandeh. "Review of swarm intelligence-based feature selection methods." *Engineering Applications of Artificial Intelligence,Apri*l 2021 doi:10.1016/ j.engappai.2021.104210.

[38] M.Rostami, S.Forouzandeh, K.Berahmand, M.Soltani. "Integration of multi-objective PSO based feature selection and node centrality for medical datasets." Genomics. November 2020;112(6):4370-4384. doi: 10.1016/j.ygeno.2020.07.027.

[39] M.Rostami., K.Berahmand. & S.Forouzandeh. "A novel community detection based genetic algorithm for feature selection." J Big Data 8, 2 January 2021. doi:10.1186/s40537-020-00398-3.

[40] M.Rostami., K.Berahmand. & S.Forouzandeh. "A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. " J Big Data 7, 83.October 2020.doi:10.1186/s40537-020-00352-3.

Ziying Song (Member, IEEE) received the B.S. degree from Hebei Normal University of Science and Technology, in 2019. He is currently pursuing a master's degree in computing technology in information Engineering and Technology at Hebei University of Science and Technology. His research interests include machine learning, computer vision, natural language processing federated learning and secure multiparty computing.



Yu Zhang received the B.S. degree from Liren College of Yanshan University, in 2019. He is currently pursuing a master's degree in computing technology in information Engineering and Technology at Hebei University of Science and Technology. His research interests include deep learning, computer vision, natural language processing and computer network.

Yi Liu received the B.S. degree from Hebei GEO University, in 2019. She is currently pursuing a master's degree in computing technology in information Engineering and Technology at Hebei University of Science and Technology. Her research interests include machine learning, computer vision and natural language processing.**Kuihe Yang** is a doctor, professor and master tutor. In 1988, he graduated from Tianjin University with a bachelor's degree; in 1997, he graduated from University of Science and Technology Beijing with a master's degree; in 2004, he graduated from Xidian University with a doctor's degree in computer application technology; in 2005, he entered Army Engineering University of PLA to do postdoctoral research in the Post-flow Station, and left the station in 2007. The main research directions are computer network, database application technology and machine learning.

Meiling Sun received the B.S. degree from Hebei University of Science and Technology, in 2020. She is currently pursuing a master's degree in computing technology in information Engineering and Technology at Hebei University of Science and Technology. Her research interests include machine learning, computer vision and natural language processing.