# Lightweight underwater object detection based on image enhancement and multi-attention

Jixiang Cheng[1*], Tian Tian[1], Dan Wu[1] and Zhidan Li[1]

[1*]School of Electrical Engineering and Information,Southwest Petroleum University, Chengdu, 610500, China.

*Corresponding author(s). E-mail(s): chengjixiang0106@126.com;
Contributing authors: 1281093756@qq.com; 2869883375@qq.com;
dan.807@163.com;

**Abstract**

Underwater object detection, which is crucial to the exploration and exploitation of marine resources, remains a challenge because noisy, weak contrast, and color distorted images are provided as sources of supervision. To address the issues of low detection accuracy caused by imprecise images, and inefficiency due to huge amount of parameters in most deep neural networks, this paper proposed a noval lightweight deep learning model with image enhancement and multi-attention. First, image enchancement algorithm MSRCR is applied to enhance image quality in order to improve the training effect of deep learning model. Then, YOLOX is used as baseline model and GhostNet is utilized as backbone network in order to reduce computation budget. Finally, a multi-attention module LCR considering level, channel and spatial domains is divised and integrated into the feature pyramid network to enhance feature learning ability and detection accuracy. Experimental result shows that on the conidered datasets our model acchieves an mAP of 77.32% and a size of 18.5MB, 1.25% higher and 46.4% less than the values of baseline network, indicating that our method achieve a superior detection precison while keeping model lighweight.

**Keywords:** underwater object detection, deep learning, lightweight model, image enhancement, attention mechanism

# 1 Introduction

With the increasing demand for raw material resources, the marine, which is rich in mineral, chemical and biological resources, has become the focus of human exploration. At present, only 5%−10% of the ocean has been exploited by human beings[26]. With the rise of artificial intelligence and robots, underwater object detection technique has been widely applied in the marine fields such as marine species exploration, marine mineral collection and marine environmental protection, which paves a fast and efficient way for marine exploration and exploitation.

Underwater object detection is a special application of general object detection, which is more challenging because of complex underwater enviroment[7]. On the one hand, the absorption and scattering of light in the underwater environment directly causes poor contrast and severe color degradation of images, and low transparency and high concentrations of microscopic inorganic and/or organic matter underwater prevents light from propagating, resulting in loss of visibility. These issues significantly lower the quality of underwater images and have a negative impact on the detection of underwater objects. On the other hand, object detection based on deep learning has achieved considerable success in computer vision tasks and have been widely applied in various scenarios. To enhance detection accuracy, most of methods use baseline network with complex structures and a large number of parameters while ignoring real-time and lightweight portability. In light of the issues in underwater object detection, research into a compact model with superior accuracy is neccesary.

At present, object detection based on deep learning can be divided into two-stage methods and one-stage methods[42]. Two-stage methods first search the region of interest selectively, then classify and regress the region. Region-based Convolutional Network (R-CNN)[13] is the pioneer network. The one-stage methods abandon the extraction of regions of interest in images but classify and regress directly after feature extraction, such as YOLO[31], SSD[24]. The one-stage methods are end-to-end approaches that simplify the process and ensures the accuracy with less computation than two-stage methods, therefore, it has stronger application potential in underwater object detection scenarios. In this work, YOLOX[10], a recently proposed excellent one-stage method, is selected as the baseline network for our study. However, the bockbone CSP-DarkNet53 with amount of parameters is utilized in YOLOX, which hinges its application to underwater object detection scenario. To achieve a high accuracy and low delay performance, image enhancement technique is applied to enchance the quality of underwater images, and a lightweight detection model with GhostNet[14] as backbone network and feature pyramid network with multi-attention module is proposed. The contributions of this are as follows:

1. Image enhancement: Multi-Scale Retinex with Color Restoration (MSRCR) algorithm is employed to enhance the color and illumination information of underwater images in order to improve detection accuracy of the deep learning model.

2. Lightweight backbone: The lightweight GhostNet is used as the backbone to replace the orignial CSPDarkNet53 so as to reduce model size and improve efficiency.
3. Fusion attention module: A fusion attension module considering scale, channel and space domains is devised and introduced into feature pyramid network to enhance feature learning ability of the model and improve detection accuracy.
4. Superior detection performance: The experimental result on three datasets shows that the proposed method gets better detection accuracy and reduce model complexity dramatically compared with existing mainstream models.

The remainder of this paper is organized as follows. In Section 2, works on object detection and lightweight network are briefly reviewed and motivation of the work is given. In Section 3, image enhancement for underwater images is presented with a preliminary analysis experiment. In Section 4, the proposed method is decribled in detail. In Section 5, experiments and analyses are carried out to demonstrate superiority of the proposed method. In Section 6, the conclusion of this paper is made.

# 2 Related works and motivation

## 2.1 Object detection

General object detection algorithms can be divided into traditional methods and deep learning based methods. The traditional methods involve amount of manual interventions and many steps. In 2001, Viola P. and Jones M.[35] proposed a sliding window method named V-J detector, which integrates the technologies of image integration, feature selection and detection cascade. In 2005, Dalal N. and Triggs B.[5] observed that the directional density distribution of gradient or edge could well describe the characteristics of local object area, and proposed Histogram of Oriented Gradient (HOG) method via using statistics of Gradient information. The method combining with support vector machine shows excellent effect. In 2008, Felzenszwalb R. *et,al.*[8] invented Deformable Part Model (DPM), showing its superiority and robustness to deformed object detection. Later, Girshick R.[11] improved DPM by introducing a cascade structure into the model, achieving more than 10 times fast speed without sacrificing accuracy. Traditional methods treat object detection as a classification problem. Through elaborately feature extraction procedures, traditional methods show good results on early limited and simple tasks. However, when facing complex scenario like underwater object detection, tradition methods fail to produce reasonable performace. Recently, due to glorious feature representive ability of deep neural networks, deep learning based methods attract wildely attention and show excellent detection performance.

Object detection based on deep learning can be classified into two-stage methods and one-stage methods. R-CNN proposed by Girshick R. *et,al.*[13] in 2014 is the pioneer work of two-stage methods. It selectively searches the

region of interest for each image, and performs feature extraction, classification and regression prediction through convolutionary neural networks. Later, the authors proposed a Fast R-CNN variant to reduce complexity of R-CNN [12], where feature extraction is conducted on the whole image. Subsequently, Faster R-CNN[32] is proposed by replacing selective search in Fast R-CNN with region proposal network and introducing the concept of anchor frame, which greatly improved the running speed. R-CNN series are representative of two-stage object detection methods with good detection accuracy. However, due to large number of model parameters, the detection speed is relatively slow. In 2015, Joseph R. *et,al.*[31] proposed the first one-stage object detection model named YOLO. The model eliminates region proposal step and uses a single neural network to directly predict boundary boxes and object categories, leading to a real end-to-end object detection network. Later, Joseph R. and Farhadi A. *et,al.*[29] released YOLOv2, which combines the training procedures of detection and classification and improves the multi-object framework. To further improve the detection performance on small objects, the authors[30] proposed YOLOv3, which uses DarkNet53 as backbone and introduces the idea of feature pyramid network to improve detection accuracy while maintaing the speed advantage. Subsequently, Bochkovskiy A. *et,al.*[2] proposed YOLOv4. The model uses CSPDarkNet53 as backbone, and spatial pyramid pooling[15] and path aggregation network[23] feature pyramid structures to further improve the performance. In 2021, Ge Z. *et,al.*[10] proposed YOLOX, which uses an anchor free idea and decoupling head for classification and regression prediction, leading to a new height for detection speed and accuracy. Besides, in 2016, Liu W. *et,al.*[24] proposed a SSD object detector that uses VGG16[34] as the backbone and abandones the boundary box recommendation and resamping procedures. The main idea is to discret output space of boundary boxes into default boundary boxes, and perform classification prediction and boundary box adjustment on the default boxes. The method achives a high detection precision when tackling low resolution images. The first-stage methods are end-to-end approaches with less parameters, and show better performance in the tasks where real-time is required.

Although showing excellent performance on general detection tasks, mainstream object detection models still face challenges in underwater scenario due to the degradation of images. Therefore, many innovative work has been proposed to tackle the challenges. For example, to solve the problem caused by object overlap, Lin W.H. *et,al.*[22] proposed ROIMIX that can well represent the interaction between images and show strong generalization ability and high accuracy in underwater object detection tasks. Li C.Y. *et,al.*[19] proposed a model named DUWIENet to enhance image quality by fusing the input and predicted confidence graph. Then the model was combined with general detection network, getting a better dectection performance on underwater images. Chen L. *et,al.*[3] proposed a Sample Weighted Hybrid Network (SWIPENet) for small object detection. By extracting high-resolution and semantically rich feature maps and using a novel sample re-weighting algorithm to reduce the

impact of noise, the method achieves superior detection accuracy. Recently, Yeh C.H. *et,al.*[39] proposed a lightweight network for jointly learning of underwater image color transformation and object detection to improve the detection performance while maitaining a low computational complexity.

## 2.2 Lightweight models

Lightweight models, aiming to reduce the number of parameters and computation complexity of networks without sacrificing the performance, has become a hot research branch in deep learning. At present, lightweight deep learning models are developing rapidly and amount of work have been proposed. In 2017, Andrew G. *et,al.*[17] proposed MobileNetV1. The central idea is replacing standard convolution with the deep separable convolution which is small in size and requires less computation. In 2018, Sandler M. *et,al.*[33] proposed MobileNetV2, which introduces linear bottleneck and reverses residual module to enable feature information to flow fully in each layer. The model implements repeated use of feature in forward propagation and alleviates the issue of gradient disappearance during back propagation. In 2019, the original team[16] devised SE module and proposed MobileNet V3, which is able to automatically acquire the importance of each feature channel through learning and improve efficiency of feature learning. In 2017, Zhang X. *et,al.*[40] proposed a lightweight model ShuffleNet V1 using pointwise group convolution and channel shuffle which dramatically reduce model complexity. Later, to meet the requirements of fewer parameters, faster speed, and higher accuracy for deploying model to mobile terminal device, Ma N. *et,al.*[25] proposed ShuffleNetV2, which uses direct measurement for evaluation and introduces channel split operation to divide feature images. In 2020, Han K. *et,al.*[14] proposed GhostNet, where the central idea is generating a few feature maps through standard convolution operation and generating new feature maps through cheap operation, and the two groups of feature maps are splintered. Experiment showed that the model achieves better performance than MoboileNet V3 within the same computation budgets. For more information about lightweigh models, refer to [41].

## 2.3 Motivation

Although many general object detection methods have been proposed, they fail to perform well in complex scenarios. In the practice of marion research, underwater object detection often requires a high real-time performance as the models need to be transplant into mobile or embedded devices. Therefore, models with suprior dectection accuracy and fewer parameters are prefered. In addition, preprocessing of images are effective ways to improve detection accuracy. Most object detection methods adopt simple image operations such as translation, rotation and mirror image to expand dataset, however, they are difficult to enhance the features learned from the network, so the effect of improving detection accuracy under complex underwater environment is

limited. To circumvent the problem, researchers propose several strategies to enhance image quality at pixel level, including transform domain methods and spatial domain methods[36]. The former ones transform the pixel and position information of original image into others space by using Fourier transform[6], wavelet transform [38], etc. The latter ones like histogram equalization[1] and Retinex[9] aim to guide the pixel space to be redistributed for better visual effect. Due to the characteristics of high blur, low color contrast and color distortion of underwater images, spatial domain methods are preferable. Based on these considerations, this paper proposes a new underwater object detection method by using image enhancement and designing a lightweight deep learning model. For image enhancement, the Retinex algorithm is applied. For model lighweight, YOLOX is as choosen as baseline and lightweight network GhostNet is used as backbone to extract features with the purpose of balancing accuracy and portability of the model. To further improve accucacy, a fusion attention module LCS is designed and introducted into the model. Experimental results show that our proposed method can not only improve detection accuracy, but also significantly reduce model complexity.

## 3 Image enhancement for underwater images

In underwater senario, the phenomenon of weak contrast and color deviation caused by light absorption and dispersion makes detection networks unable to efficiently learn the feature information of underwater objects. Therefore, before been feeding into network, the images could be first enhanced. In this work, Retinex [9] is adopted, which was devised based on three principles. First, there is no color distinction in real world, and the color perceived by human is the result of interaction between objects and light. Second, colors are composed of three primary colors red, green and blue. Third, the three primary colors determine the color of units in different areas. It believes that the color is determined by light reflection ability of three primary colors and has nothing to do with light intensity. The algorithm simulates the relationship between human retina and cerebral corte, and imitates human visual system to achieve color balances. Retinex works as follows. When an image is illuminated on the reflected object by incident light, the image $s(x, y)$ seen by observer is composed of incident image $l(x, y)$ and reflected image $r(x, y)$, expressed as

$$s(x, y) = r(x, y) \cdot l(x, y) \tag{1}$$

To imitate the process of human perception of images, the method applies logarithmic oprerator to the $r(x, y)$, resulting

$$\log s(x, y) = \log r(x, y) + \log l(x, y) \tag{2}$$

Denoting $S = \log s(x, y)$ , $R = \log r(x, y)$ ,$L = \log l(x, y)$ , then the incident image can be obtained by

$$L = S - R \tag{3}$$

where $L$ is the enhanced image after restoration.

Since Retinex beening devised, servaral variants have been proposed, and some well-known ones are Multi-Scale Retinex (MSR) [28], MSR with Color Restoration (MSRCR)[27], and MSR with Chromaticity Preservation (MSRCP)[21] .

MSR is a multiple Retinex algorithms with different weights. Its principle is formulated as

$$F_{\mathrm{MSR}}(x,y)=\sum_{j=1}^{N}\omega_j\{\log s_j(x,y)-\log[f_j(x,y)*s_j(x,y)]\} \tag{4}$$

where, $F_{\mathrm{MSR}}$ is the enhancement result, * is the convolution operation, $\omega_j$ is the weight factor satifying $\sum\omega_j=1$, and $f_j(x,y)$ is the filtering function formulated as

$$f_j(x,y)=k\cdot\exp(-\frac{x^2+y^2}{2{\sigma_j}^2}) \tag{5}$$

where $k$ is a key parameter determining final enhancement effect satisfying $\iint f(x,y)\mathrm{d}x\mathrm{d}y=1$ .

To correct chromatic aberration and restore the true color of images, MSRCR introduces color restoration factor to MSR, working as

$$F_{\mathrm{MSRCR}}(x,y)=c_j(x,y)\cdot F_{\mathrm{MSR}}(x,y) \tag{6}$$

where, $F_{\mathrm{MSRCR}}$ is the enhancement result, $c_j(x,y)$ is the color recovery factor calculated by

$$c_j(x,y)=\mu\cdot\log[\eta\cdot\frac{s_j(x,y)}{\sum\limits_j s_j(x,y)}] \tag{7}$$

where $\mu$ and $\eta$ are constants, representing color recovery gain factor and offset quantity, respectively.

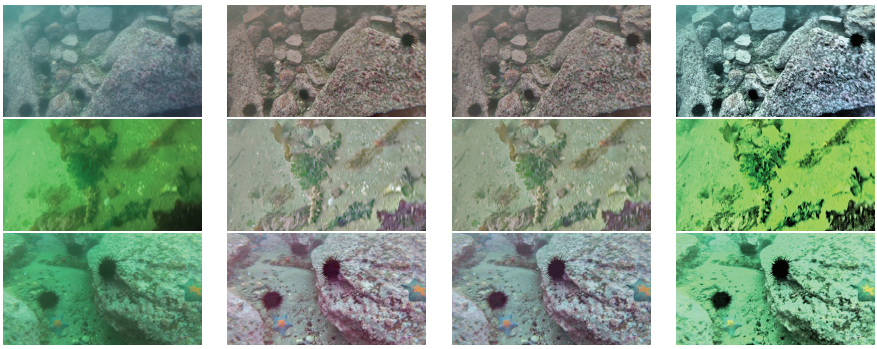
MSRCP adds color amplification factor to MSR with the purpose of enhancing the image while retaining color distribution. The principle is formulated as

$$F_{\mathrm{MSRCR}}(x,y)=F_{\mathrm{MSR}}(x,y)\cdot\max[I_R(x,y),I_G(x,y),I_B(x,y)] \tag{8}$$

where $I_R$ , $I_G$ and $I_B$ are the original RGB channels, respectively.

MSR, MSRCR and MSRCP are applied to enhance fuzzy, greenish and bluish underwater images from URPC2021 dataset with four species as examples. Fig. 1 illustrates the enhancement results of three randomly choosen underwater images. According to the result, MSR and MSRCR show better color correction effect, but the defogging effect is still insufficient, and MSRCR is much suitable for partial blueprint images. MSRCP could enhances the contrast of blurred images, however, it has little effect on correcting color deviation.

To further inverstigate the effects of three image enhancement algorithms on underwater object detection, YOLOX was chosen as detection network. The orignal dataset and three enhanced datasets are applied to train and test model independelty, and the results are listed in Table 1. According to the result, both MSR and MSRCP enhances detection accuracy, while MSRCP deteriors the result. Considering that MSRCR impoves the results most, it is choosen as the final image enhancement algorithm for our method.



(a) Original image (b) Enhanced by MSR (c) Enhanced by MSRCR (d) Enhanced by MSRCP

**Fig. 1** Comparisons of three iamge enhancement algorithms on sample images

**Table 1** Comparions of detection accuracy with different image enhancement algorithms

| | AP(%) | | | | |
|---|---|---|---|---|---|
| Enhancement algorithm | echinus | starfish | scallop | holothurian | mAP(%) |
| Original | 88.58 | 80.90 | 72.51 | 62.28 | 76.07 |
| MSR | 88.25 | 79.92 | **74.49** | 65.25 | 76.98 |
| MSRCR | **88.73** | **81.06** | 74.07 | **65.39** | **77.31** |
| MSRCP | 87.34 | 79.14 | 71.32 | 61.15 | 74.71 |

# 4 Proposed lightweight detection model

To reduce model size of YOLOX and enhance object detection accuracy for underwater images, a new model is proposed. The characteristics of our model are two sides. First, the lightweight network GhostNet is used as backbone network to reduce the number of weights. The second is a LCS module of fusion attention mechanism is devised and integrated into the model to imporve the accuracy. The overall structure of our proposed model is illustrated in Fig. 2.

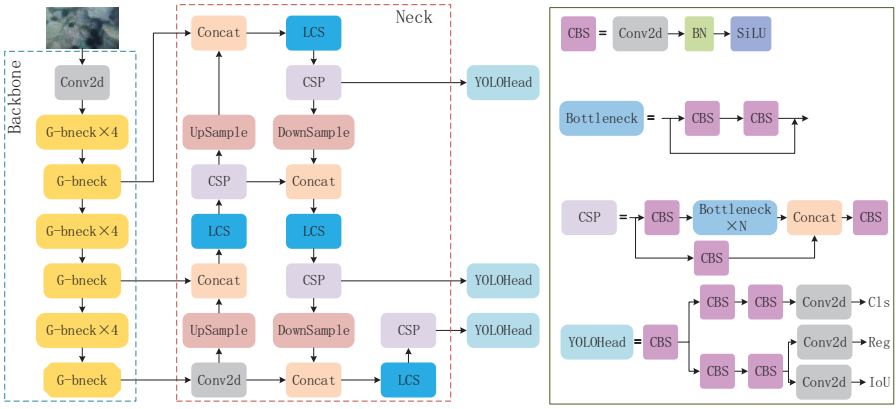


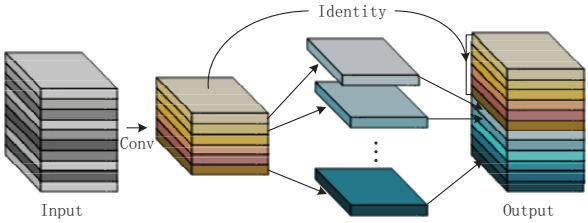

**Fig. 2** Overall structure of our proposed model

## 4.1 Lightweight backbone network

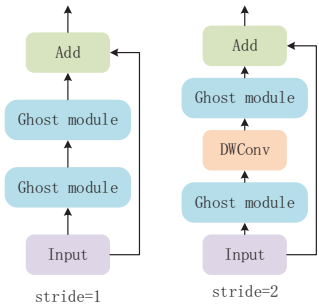
The feature extraction mainly depends on the backbone network. The CSP-Darknet53 of YOLOX contains too many parameters, which makes the network complex. To circumvent the issue, GhostNet is adopted in our model as backbone network. The GhostNet is mainly composed of Ghost bottleneck (G-bneck), which uses Ghost convolution module to replace normal convolution in residual structure. There are two types of G-bneck, as shown in Fig. 3. When only channel transformation is carried out, the stride size is set as 1. When the compression of feature maps is required, the stride size is set to 2, and a depth-separable convolution operation is added between two Ghost modules. For our task, the images are first cropped into the size of 640×640×3, and then a 3×3 convolution with strip size 2 is applied. Afterwards, 16 G-Bnecks are used to extract depth features. To ensure that the extracted features contain rich semantic and location information and make multi-scale features fully integrated, three feature layers with size of 80×80×40, 40×40×112 and 20×20×160 are selected as the input of neck network.

## 4.2 LCS attention module

Pixel features with the same labels are easily affected by local receptive fields during convolution operation, which may lead to inconsistent classification results and it is much prominent for small objects. To solve the problem, a multi-attention mechanism module LCS is divised to make the network focus on the context information of objects with different sizes so as to obtain better pixel-level feature representation. The structure of LCS module is shown in Fig. 4. The module is composed of level attention, channel attetion and spatial attension, so that it enables the model learning object features and corresponding positions better.




(a) Ghost module




(b) G-bneck

**Fig. 3** Structure of Ghost module and G-bneck

Level Attention (LA)[4] is fused according to the semantic importance of different scale features to enhance the scale perception ability of the network. First, channel features of input feature maps are processed by average pooling, then the sum of channel pixels is obtained by convolution operation. Finally, the activation function is used to obtain scale attention weight which is multiplied by original feature maps. The calculation can be expressed as

$$OP_{\mathrm{LA}} = IP \cdot \delta(\mathrm{Conv}(\mathrm{AvgPool}(IP))) \tag{9}$$

where $OP_{\mathrm{LA}} \in R^{H \times W \times C}$ is the output feature maps, $IP \in R^{H \times W \times C}$ is the input feature maps, AvgPool is the average pooling operation, Conv is the convolution operation, $\delta$ is Sigmoid activation function, respectively. The size of scale attention weight parameter $\delta(\mathrm{Conv}(\mathrm{AvgPool}(IP)))$ is $1 \times 1 \times 1$.

Channel Attention (CA)[18] enhances feature discrimination by establishing semantic dependence between feature maps. Firstly, channel features of input feature maps are processed by maximum pooling function. Then linear maps are obtained through a full connection layer. Finally, the activation function is used to obtain channel attention weight which is multiplied by original feature maps. The calculation can be expressed as

$$OP_{\mathrm{CA}} = IP \cdot \delta(\mathrm{FC}(\mathrm{MaxPool}(IP))) \tag{10}$$

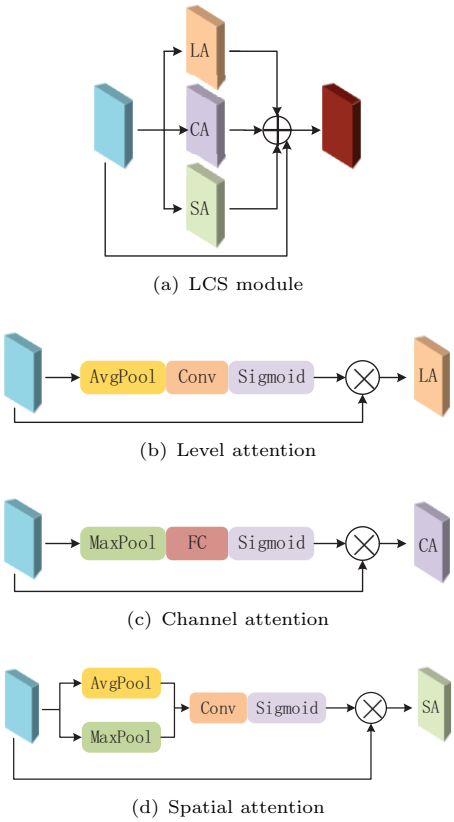



**Fig. 4** LCS module structure

Where $OP_{\mathrm{CA}} \in R^{H\times W\times C}$ is the output feature maps with channel attention, MaxPool is the maximum pooling operation, FC is a full connection layer, respectively. The size of channel attention weight parameter $\delta(\mathrm{FC}(\mathrm{MaxPool}(IP)))$ is $1 \times 1 \times C$.

Spatial Attention (SA)[37] exploits the relationship between features to generate a spatial feature map, so that the network could pay more attention to the location information with strengthening learning of the regions of interest while weakening other regions. First, maximum and average pooling are conducted in parrallel on the input feature maps. Then, a 7×7 convolution operation is used to extract features and reduce the number of channels. Finally, sigmpoid activation function is performed to obtain spatial attention weight parameters. The calculation can be expressed as

$$OP_{\mathrm{SA}} = IP \cdot \delta(\mathrm{Conv}(\mathrm{MaxPool}(IP), \mathrm{AvgPool}(IP))) \tag{11}$$

Where $OP_{\mathrm{SA}} \in R^{H\times W\times C}$ is the output feature. The size of spatial attention weight parameter is $H \times W \times 1$.

By integrating the LCS module, the network can obtain better global context information and is able to pay more attention to the feature changes from scale, channel and spaital aspects. Therefore, the model can effectively solve the problem of low detection accuracy caused by the feature difference of the same objects.

# 5 Experimental results

To verify the effectiveness and generalization of the proposed model, we perform a comprehensive quantitative comparison with the state-of-the-art algorithms on real world images. Methods used for comparison include SSD[24], YOLOV3[30], YOLOV4[2], NanoDet[20] and YOLOX[10]. To make a fair comparison, both the proposed method and those methods used for comparison were carried out under the same experimental environment.

## 5.1 Datasets

Three datasets, i.e., URPC2021, Brackish and VOC2007, are employed to conduct experiments. The URPC2021 dataset, from underwater robot optical image competition, contains 7543 labled underwater optical images with four spieces, including starfish, sea urchins, sea cucumbers and scallops. The images have features of weak color, low contrast and object overlapping, making model generalizes hard. The Brackish dataset is a public real underwater images, derived from a video shot in a Brackish channel. The dataset contains six species, including fish, minnow, crab, shrimp, jellyfish and starfish, with features of blurring and color deviation. The VOC2007 dataset is a general dataset in object detection research area, which contains 5011 labeled images of twenty kinds of objects. The VOC2007 dataset is choosen to verify the generalization performance of our model.

## 5.2 Experiment setup

The stratified sampling method was adopted to divide the dataset to avoid the risk of sampling bias. Random seeds were set during the division of training and test samples so that the proposed method and the compared methods used the same training samples. For all datasets, the training proportion, validation proportion, and test proportions were set to 80%, 10%, and 10%, respectively. The equipment configuration and hyper-parameters of proposed mathod are listed in Table 2. For comparison methods, the hyper-parameter settings in their orignal papers are applied.

## 5.3 Experimental Results

### 5.3.1 Quantitative comprisons on underwater datasets

To verify the effectiveness of the proposed method on underwater datasets, URPC2021 and Brackish are considered. The experimental results on

**Table 2** Experimental environment and parameter settings

| Item | Contents |
|---|---|
| Processor | Inteli7-9700K@3.6GHz |
| Memory | 32GB |
| Operating system | Ubuntu16.04LTS |
| Solid state hard disk | 1TB |
| GPU | NVIDIA Geforce RTX2080Ti |
| Pytorch | v1.10 |
| Learning rate | 0.0001 and Cosine annealing function is used |
| Freeze epoch | 50 |
| Unfreeze epoch | 150 |
| Batch size | 4 |
| Image size | 640 × 640 |

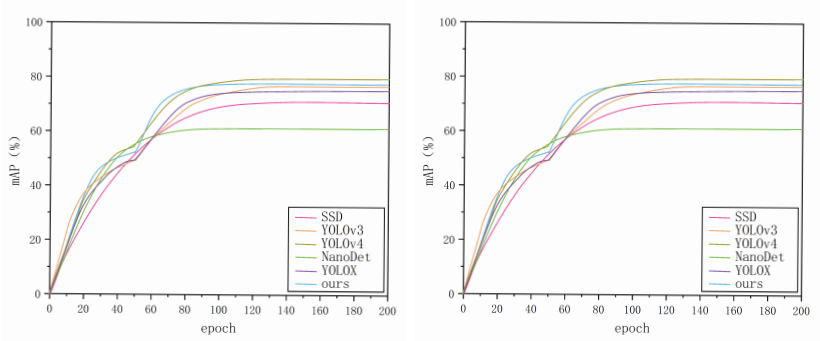
URPC2001 are listed in Table 3. It can be seen that the detection accuracy of the proposed method is superior to SSD and YOLOv3, and the model size is only 20% and 7.9% of the two ones, respectively. YOLOV4 achieves the highest overall accuracy, but with a large number of parameters, while the size of our model is only 7.6% of YOLOV4 . Althouth NanoDet gets the highest detection speed, the average detection accuracy is quite worse and cannot be adapted to object detection in underwater scenarios. Due to the relative high quality of images in Brackish dataset, image enhancement is not applied for our method. The comparive results are listed in Table 4. Although the average accuracy of proposed method is slightly inferior that those of YOLOv4 and YOLOv3, the model size is reduced dramatically and detection speed is enhanced to a large degree. NanoDet has the lowest model weight and the highest FPS, but the average accuracy is too low. Compared with SSD and YOLOX, the proposed method is superior in terms of average accuracy, model weight and FPS value. Fig. 5 gives the mAP cruves on two datasets. The curves show that the model tends to be stable at about 100th and 120th epoch, indicating a fast learning efficiency and strong stability. Therefore, our proposed method can achieve excellent detection accuracy while keep the model slim and adapt well in complex underwater object detection scenario.

**Table 3** Comparative results on URPC2021 dataset

| | AP(%) | | | | | | |
|---|---|---|---|---|---|---|---|
| Algorithm | echinus | starfish | scallop | holothurian | mAP(%) | Model size(MB) | FPS |
| SSD | 84.33 | 79.03 | 56.42 | 66.50 | 71.57 | 92.1 | 41.77 |
| YOLOv3 | 86.69 | 78.91 | 76.06 | 66.83 | 77.12 | 235.3 | 34.86 |
| YOLOv4 | **89.27** | 82.19 | **77.97** | **69.80** | **79.81** | 244.7 | 37.55 |
| NanoDet | 77.72 | 63.56 | 43.98 | 60.44 | 61.34 | **17.1** | **60.32** |
| YOLOX | 88.58 | 80.90 | 72.51 | 62.28 | 76.07 | 34.3 | 47.01 |
| Ours | 87.86 | **85.83** | 73.27 | 62.33 | 77.32 | 18.5 | 55.54 |

**Table 4** Comparative results on Brackish dataset

| | AP(%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | Fish | Small fish | Crab | Shrimp | Jellyfish | Starfish | mAP(%) | Model size(MB) | FPS |
| SSD | 89.47 | 76.59 | 82.41 | 88.44 | 87.29 | 90.64 | 85.81 | 91.9 | 42.31 |
| YOLOv3 | 95.72 | 80.31 | 90.45 | 92.66 | 93.28 | 93.16 | 90.93 | 234.7 | 35.88 |
| YOLOv4 | **97.45** | **81.88** | **91.79** | **94.21** | **95.07** | 95.75 | **92.69** | 244.0 | 38.74 |
| NanoDet | 87.49 | 72.19 | 74.52 | 81.45 | 81.33 | 80.98 | 79.66 | **16.8** | **61.98** |
| YOLOX | 95.06 | 80.53 | 87.27 | 91.44 | 91,18 | 93.76 | 89.87 | 34.0 | 47.56 |
| Ours | 94.48 | 80.31 | 89.58 | 92.17 | 92.62 | **95.87** | 90.84 | 18.3 | 55.27 |




(a) UPRC2001 dataset (b) Brackish dataset

**Fig. 5** Comparisons of mAP-Epoch curves

### 5.3.2 Underwater detection effect

Fig. 6 illustrates some underwater object detection effect of all the methods on three samples images, where the left two samples are randomly selected from UPRC2001 dataset and the third one is selected from Brackish. The detected objects are marked with rectangular boxes and corresponding categories. It can be seen that the proposed method is able to well detect the objects with shade background and small object with overlap. Therefore, the proposed method perform well in detecting fuzzy, overlapping and occluded objects, reflecting a superior overall performance.

### 5.3.3 Ablation experiment

To inverstigate the effectiveness of the components proposed in our method, URPC2021 dataset is utilited to conduct an ablation experiment. The results are listed in Table 5, where YOLOX is the baseline. When GhostNet is used as

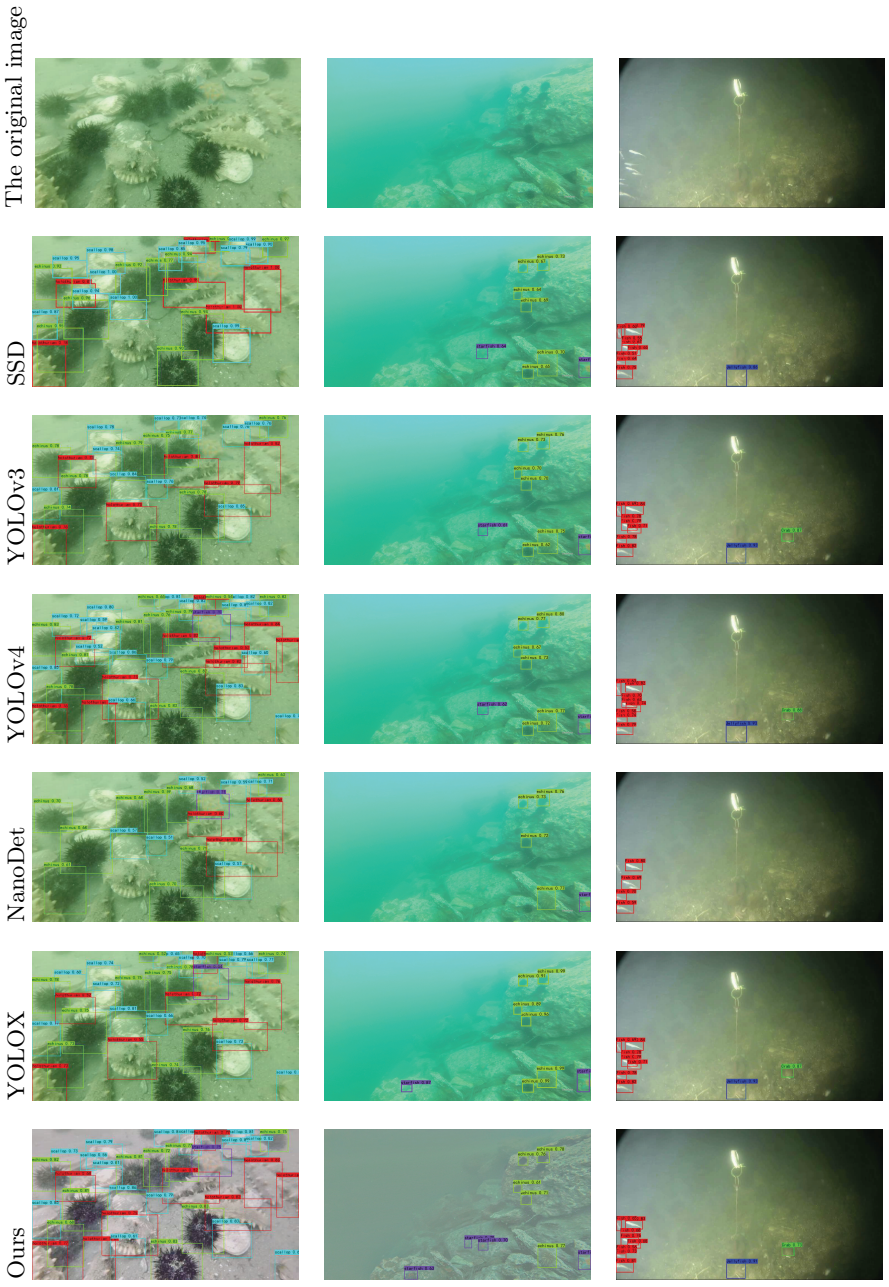



**Fig. 6** Detection performance comparions of all the algorithms on sample images

backbone, the weight is reduced by 46.4% accompanied with an accuracy loss of 1.02%. When LCS module is integrated into modified backbone network, the detection accuracy increases by 1.276%, compensating the accuracy loss due to the backbone change. Meanwhile, the model size is increased by only 0.1MB, which indicates that the proposed LCS module is able to strengthen the feature learning ability while not sacrifising complexity. When MSRCR is performed, the detection accuracy is further increases by 1%, indicating that image enhancement is able to strengthen object feature information and helpful to enhance training effect. Compared with baseline, the accuracy of the proposed method is improved by 1.256% with the weight reducing by 46.46%, achieving the goal of improving detection accuracy and realizing model lightweight simutaneously. In addition, the PR curve of our method is shown in Fig. 7, indicating that the method gets higer detection accurcy on sea urchins and starfish while works worsely on scallops and sea cucumbers. Hence, to further enhance the accurcy, more effort could be paid to such two categories.

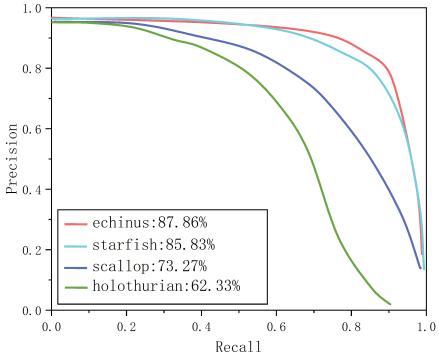



**Fig. 7** PR curve of the proposed method

**Table 5** Results of ablation experiment

| method | | | | AP% | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ① | ② | ③ | ④ | echinus | starfish | scallop | holothurian | mAP(%) | Model size(MB) |
| ✓ | | | | 88.58 | 80.90 | 72.51 | 62.28 | 76.07 | 34.3 |
| | ✓ | | | **89.45** | 80.03 | **73.93** | 56.80 | 75.05 | **18.4** |
| | ✓ | ✓ | | 88.28 | 84.28 | 72.25 | 60.45 | 76.32 | 18.5 |
| | ✓ | ✓ | ✓ | 87.86 | **85.83** | 73.27 | **62.33** | **77.32** | 18.5 |

①: Baseline, ②: Lightweight backbone, ③: LCS module, ④: MSRCR.

### 5.3.4 Comparison experiment of general dataset

To verify generalization performance of the proposed method without image enchancement, further experimemt is carried out on general VOC2007 dataset and the results are listed in Table 6. The results show amost the same performence as on underwater dataset. The average accuracy is only second to YOLOv4 with a dramatically model size reduction. Although the size and detection speed is slightly worse than NanoDet, the accuracy is much better. Therefore, it can be concluded that the proposed method not only achieves a good balance between accuracy and speed, but also shows a strong generalization performance in other scenarios.

**Table 6** Comparative results on VOC dataset

| Algorithm | mAP(%) | Model size(MB) | FPS |
|---|---|---|---|
| SSD | 76.55 | 92.7 | 47.32 |
| YOLOv3 | 85.71 | 235.8 | 34.93 |
| YOLOv4 | **89.93** | 244.9 | 38.57 |
| NanoDet | 72.58 | **17.1** | **62.74** |
| YOLOX | 86.07 | 34.6 | 48.04 |
| Ours | 86.68 | 19.1 | 57.83 |

# 6 Conclusion

In this paper, we present a lightweight deep learning model for underwater object dection that combiness the advantages of image enhancement, lightweight backbone and attention mechanism. MSRCR algorithm is applied to enhance the quality of underwater images with weak contrast and color distoration in order to improve the learning effect. GhostNet is used as the backbone to dramatically reduce the number of parameters to implement lightweight. A fusion of attention module with level attention, channel attetion and spatial attention is devised to extract more meaningful features in order to enchance the dection accuracy. Experiments were carried out on three datesets. The results demonstrate that the proposed method makes better balance on detection accuracy and complexity than the state-of-the-art methods on complex underwater object dection task, and shows a strong generalization performance on general detection task. Recently, model compression and transformer have attracted widely attentions and achieved remarkable results on various tasks. In future work, we will inverstigate their applications on underwater object detection in order to further enhance the model performance.

# References

[1] Abdullah-Al-Wadud M, Kabir MH, Akber Dewan MA, et al (2007) A dynamic histogram equalization for image contrast enhancement. IEEE Transactions on Consumer Electronics 53(2):593–600. https://doi.org/10.1109/TCE.2007.381734

[2] Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv:200410934

[3] Chen L, Liu ZH, Tong L, et al (2020) Underwater object detection using invert multi-class adaboost with deep learning. In: International Joint Conference on Neural Networks (IJCNN), pp 1–8, https://doi.org/10.1109/IJCNN48605.2020.9207506

[4] Dai X, Chen Y, Xiao B, et al (2021) Dynamic head: Unifying object detection heads with attentions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7369–7378, https://doi.org/10.1109/CVPR46437.2021.00729

[5] Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 886–893, https://doi.org/10.1109/CVPR.2005.177

[6] Ell TA, Sangwine SJ (2007) Hypercomplex fourier transforms of color images. IEEE Transactions on Image Processing 16(1):22–35. https://doi.org/10.1109/TIP.2006.884955

[7] Fayaz S, Shabir AParah, Qureshi G (2022) Underwater object detection: architectures and algorithms – a comprehensive review. Multimedia Tools and Applications 81:20,871–20,916. https://doi.org/10.1007/s11042-022-12502-1

[8] Felzenszwalb PF, Girshick RB, McAllester D, et al (2010) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1627–1645. https://doi.org/10.1109/TPAMI.2009.167

[9] Fu X, Zhuang P, Huang Y, et al (2014) A retinex-based enhancing approach for single underwater image. In: IEEE International Conference on Image Processing (ICIP), pp 4572–4576, https://doi.org/10.1109/ICIP.2014.7025927

[10] Ge Z, Liu ST, Wang F, et al (2021) Yolox: Exceeding yolo series in 2021. arXiv:210708430

[11] Girshick R (2012) From rigid templates to grammars: Object detection with structured models. PhD thesis, USA

[12] Girshick R (2015) Fast r-cnn. In: IEEE International Conference on Computer Vision (ICCV), pp 1440–1448, https://doi.org/10.1109/ICCV.2015.169

[13] Girshick R, Donahue J, Darrell T, et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 580–587, https://doi.org/10.1109/CVPR.2014.81

[14] Han K, Wang YH, Tian Q, et al (2020) Ghostnet: More features from cheap operations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1577–1586, https://doi.org/10.1109/CVPR42600.2020.00165

[15] He K, Zhang X, Ren S, et al (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9):1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

[16] Howard A, Sandler M, Chen B, et al (2019) Searching for mobilenetv3. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp 1314–1324, https://doi.org/10.1109/ICCV.2019.00140

[17] Howard AG, Zhu M, Chen B, et al (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:170404861

[18] Hu J, Shen L, Albanie S, et al (2020) Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

[19] Li CY, Guo CL, Ren WQ, et al (2020) An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing 29:4376–4389. https://doi.org/10.1109/TIP.2019.2955241

[20] Li X, Lv CQ, Wang WH, et al (2022) Generalized focal loss: Towards efficient representation learning for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–14. https://doi.org/10.1109/TPAMI.2022.3180392

[21] Lin J, Miao ZJ (2016) Research on the illumination robust of target recognition. In: IEEE International Conference on Signal Processing (ICSP), pp 811–814, https://doi.org/10.1109/ICSP.2016.7877943

[22] Lin WH, Zhong JX, Liu S, et al (2020) Roimix: Proposal-fusion among multiple images for underwater object detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2588–2592, https://doi.org/10.1109/ICASSP40776.2020.9053829

[23] Liu S, Qi L, Qin H, et al (2018) Path aggregation network for instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8759–8768, https://doi.org/10.1109/CVPR.2018.00913

[24] Liu W, Dragomir A, Dumitru E, et al (2016) Ssd: Single shot multibox detector. In: European Conference on Computer Vision (ECCV), pp 21–37, https://doi.org/10.1007/978-3-319-46448-0_2

[25] Ma NN, Zhang XY, Zheng HT (2018) Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: European Conference on Computer Vision (ECCV), pp 122–138, https://doi.org/10.1007/978-3-030-01264-9_8

[26] Miloslavich P, Seeyave S, Muller-Karger F, et al (2019) Challenges for global ocean observation: the need for increased human capacity. Journal of Operational Oceanography 12(sup2):S137–S156. https://doi.org/10.1080/1755876X.2018.1526463

[27] Parthasarathy S, Sankaran P (2012) An automated multi scale retinex with color restoration for image enhancement. In: National Conference on Communications (NCC), pp 1–5, https://doi.org/10.1109/NCC.2012.6176791

[28] Rahman Z, Jobson D, Woodell G (1996) Multi-scale retinex for color image enhancement. In: IEEE International Conference on Image Processing (ICIP), pp 1003–1006, https://doi.org/10.1109/ICIP.1996.560995

[29] Redmon J, Farhadi A (2017) Yolo9000: Better, faster, stronger. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6517–6525, https://doi.org/10.1109/CVPR.2017.690

[30] Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:180402767

[31] Redmon J, Divvala S, Girshick R, et al (2016) You only look once: Unified, real-time object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 779–788, https://doi.org/10.1109/CVPR.2016.91

[32] Ren S, He K, Girshick R, et al (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on

Pattern Analysis and Machine Intelligence 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[33] Sandler M, Howard A, Zhu M, et al (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4510–4520, https://doi.org/10.1109/CVPR.2018.00474

[34] Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:14091556

[35] Viola P, Jones M (2004) Robust real-time face detection. In: International Journal of Computer Vision (IJCV), pp 137–154

[36] Wang Y, Song W, Fortino G, et al (2019) An experimental-based review of image enhancement and image restoration methods for underwater imaging. IEEE Access 7:140,233–140,251. https://doi.org/10.1109/ACCESS.2019.2932130

[37] Woo SY, Park J, Lee JY, et al (2018) Cbam: Convolutional block attention module. In: European Conference on Computer Vision (ECCV), pp 3–19, https://doi.org/10.1007/978-3-030-01234-2_1

[38] Xu XJ, Wang YR, Yang GS, et al (2016) Image enhancement method based on fractional wavelet transform. In: IEEE International Conference on Signal and Image Processing (ICSIP), pp 194–197, https://doi.org/10.1109/SIPROCESS.2016.7888251

[39] Yeh CH, Lin CH, Kang LW, et al (2021) Lightweight deep neural network for joint learning of underwater object detection and color conversion. IEEE Transactions on Neural Networks and Learning Systems pp 1–15. https://doi.org/10.1109/TNNLS.2021.3072414

[40] Zhang XY, Zhou XY, Lin MX, et al (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6848–6856, https://doi.org/10.1109/CVPR.2018.00716

[41] Zhou Y, Chen SC, Wang YM, et al (2020) Review of research on lightweight convolutional neural networks. In: IEEE Information Technology and Mechatronics Engineering Conference (ITOEC), pp 1713–1720, https://doi.org/10.1109/ITOEC49072.2020.9141847

[42] Zou ZX, Shi ZW, Guo YH, et al (2019) Object detection in 20 years: A survey. arXiv:190505055