Journal of Computers Vol. 34 No. 2, April 2023, pp. 113-127
doi: 10.53106/199115992023043402009

Research on Rip Currents Detection Method Based on Improved YOLOv5s

Rui Qi[1], Dao-Heng Zhu[2], Xue Qin[1*]

[1] College of Big Data & Information Engineering, Guizhou University,
Guiyang 550025, China
qrlib7@163.com, xqin@gzu.edu.cn

[2] School of Electronics and Information Engineering, Guangdong Ocean University,
Zhangjiang 524088, China
dhzhu911@163.com

Received 7 May 2022; Revised 16 September 2022; Accepted 28 October 2022

Abstract. Rip currents are common natural disaster and widely distributed on beaches around the world, which can quickly bring swimmers into deep water and cause safety accidents. Rip currents are generally sudden and insidious, making it difficult for inexperienced beach managers and tourists to identify them, and presenting a high risk to swimmers. Deep learning is a popular technology in the field of computer vision, but its applications in rip currents recognition are rare, and it is difficult to realize real-time detection of rip currents. In response to the above problems, we propose an improved YOLOv5s rip currents identification method. Firstly, a joint dilated convolution module is designed to expand the receptive field, which not only improves the utilization of feature information, but also effectively reduces the amount of parameters. Then, a parameter-free attention mechanism module is added, which does not increase the complexity of the model and can improve the detection accuracy at the same time. Finally, the Neck area of the original YOLOv5s model is simplified, the 80x80 feature map branch suitable for detecting small targets is deleted, and the overall complexity of the model is reduced by reducing the amount of parameters to improve the real-time detection. We have conducted multiple sets of experiments on public data set. The results show that compared with the original YOLOv5s model, the mAP of the improved model for identifying rip currents on the same data sets has increased by 4%, reaching 92.15%, and the frame rate has increased 2.18 frames per second, and the model size is only increased by 0.45 MB. Compared with several mainstream models, the improved model not only has a simplified structure but also significantly improves the detection accuracy, indicating that our model has the accuracy and efficiency in detecting rip currents, and can provide an effective way for embedded devices to perform accurate target detection.

Keywords: rip currents, target detection, deep learning, joint expansion convolution, parameter-free attention

1 Introduction

Rip currents were first proposed by Shepard in 1936 [1], which mean that after the waves wash up on the beach, due to the uneven regional slowdown, the place where the slowdown is fast becomes the main return point. When the seawater flows back to the ocean through this point, it forms a strip-shaped, beam-shaped, fast-moving narrow and strong current [2], which is concealed, sudden, unpredictable, fast, and the direction is almost perpendicular to the coast, etc. Its speed can reach up to 3m/s, much higher than most people's swimming speed. The length can reach 30 to 100 meters or even longer [3]. Every summer, beaches can be a popular way of vacation, but rip currents occur on beaches all over the world. Most of the time, they are hidden, so that the swimmers do not notice. Once a rip current is fast, it will pull the swimmers away from the beach and quickly take them to the deep water area. The swimmers instinctively swim back, but it is difficult to reach the speed faster than the rip current, resulting in drowning death. Therefore, rip current is the leading cause of drowning in offshore beaches, also the most serious safety risk for beaches and large lake swimmers, and is a common natural disaster. On August 14, 2021, 17 tourists from Zhangzhou, a city in Fujian province of China, were swept into the sea by the rip currents, directly killing 11 people.

Due to the hidden nature of rip currents, related research has shown that only 44.5% of beachgoers believe the signs posted on the coast can help them actually identify rip currents, and among those who are confident that they can identify rip currents, the success rate of accurate identification is less than 20% [4]. According to "China

* Corresponding Author

Marine Economy Statistical Bulletin", released in March 2021, coastal tourism accounted for 47.0% of the added value of the domestic marine economy. However, there are a large number of rip currents and drowning records in some popular seaside tourist beaches, but the public and beach management departments don't have a clear concept about rip currents. Compared with the expensive and impractical methods of measuring current information with Acoustic-Doppler Velocimeters or Lowered ADCP [5], it is very urgent and necessary to research how to detect rip currents efficiently and lightly, warn beachgoers and fishing boats to stay away from dangerous areas, reduce losses and improve coastal safety.

At present, research on object detection based on deep convolutional neural network (CNN) in the marine engineering field has achieved many important results, such as wave breaking classification in infrared imagery [6], CNN was applied to estimate wave breaking type from close-range monochrome infrared imagery of the surf zone. In [7] an image processing procedure was described that includes preprocessing, segmentation, classification, and postprocessing for the accurate identification of 108 classes of plankton using spatially sparse CNNs. So rip current detection has also begun to use deep learning.

However, there are few research results on rip current detection based on deep convolution neural network at present, because the target detection technology is mainly applied to physical targets with regular trajectories, while in the marine field, rip current has irregular amorphous structure, and there is no clear contour boundary even when it can be observed. Therefore, how to use deep convolution neural network to detect rip current with high accuracy is a great challenge.

Secondly, due to the high risk of the rip current, it is more important to detect it in real-time, so as to warn the coastal tourists and ships timely. From the perspective of real-time performance, the lightweight network model can be used to reduce the amount of calculation and improve the network processing speed. MobileNet [8] replaced the traditional convolution with depthwise separable convolution, which was proposed by Google and has realized a variety of applications in mobile terminals. Subsequently, more lightweight networks with small size appeared, including EfficientDet [9], SqueezeNet [10], MobileVit [11] and so on. The single-stage network model YOLO series has also launched a tiny lightweight series, and has also achieved good research results in real-time target detection [12-15]. The detection rate is much higher than that of the two-stage network model R-CNN [16], Fast R-CNN [17], Fast R-CNN [18], etc.

Based on the above discussions, we will study how to use deep convolutional neural network to identify rip current, improve the identification accuracy, and make the model meet the lightweight requirements necessary for real-time detection, so we select YOLOv5s lightweight network, and proposes some improved rip currents identification methods of YOLOv5s.The main technical contributions are summarized as follows:

We design and add a JDC (Joint Dilated Convolution) module to the horizontal connection of the FPN (Feature Pyramid Network) [19], which will effectively reduce the amount of parameters and keep the model compact while improving the utilization of feature information. Besides this, three additional JDC structures were designed and compared experimentally to verify the validity of the original one.

SimAM (A Simple, Parameter-Free Attention Module) [20] is added to enhance the features. Let the network focus more on filtering out the high-value feature information of the rip currents from the input images, suppress other useless information, maintaining the parameter quantity unchanged and maintaining the model lightweight while enhancing the model detection ability.

considering the rip currents are large targets in the images, we remove the 80 x 80 feature map detection branch for small targets to reduce the complexity of the model and improve the detection speed of the model.

(4) The improved model was trained and tested on the rip currents data set published by Akila et al [21]. And the experimental results show that methods used in this research have a certain improvement in detection of the speed and accuracy. It can satisfy embedded devices to efficiently detect and identify rip currents based on video datas, and can also be used as the preliminary technical preparations for real-time coastal observation systems.

2 Related Work

In research of exploring visualization and identification of rip currents, at first, the simulation method is used, for example, the fluorescein dye is dumped in the area where the rip current is formed to observe the size, shape and dispersion of the rip current [22-23], or equipped with professional detector, including wave sensor, sound velocity meter or current profiler, is deployed at a specific location to detect the current velocity [24]

In [25], Floating drifters with embedded GPS units was developed to measure currents, that provides a relevant framework to understand the primary morphological and hydrodynamic parameters controlling surf-zone rip

current occurrence and dynamics. Three broad categories of rip current types are described based on the dominant controlling forcing mechanism.

The above-mentioned traditional detection of rip current is too costly, time-consuming, and low in accuracy. The detection speed cannot be guaranteed, which seriously affects the real-time performance of issuing rip current warnings to beach tourists. In contrast, integrating some special processing methods will have a very important improvement effect on detection.

In [26], Shweta et al combined the rip currents detection method with the color wheel to extract the direction of rip currents and assign corresponding colors. By setting the speed threshold, it was possible to identify areas with strong countercurrents as rip currents. But there are some limitations, as it only has a good processing effect on ocean current images with strong texture, which was only based on certain requirements for the data sets.

In [27], Yuli Liu et al developed an integrated nowcast forecast operational system to provide real-time flash rip warnings to beach users. This is a high performance and distributed computing infrastructure, digitally detects and assesses flash rip hazards in high, moderate or low risks, named "Lifeguarding Operational Camera Kiosk System (LOCKS)". The system can map the RGB image of the original rip currents to HSV space, segmented the image into regions with different morphological characteristics, and determined the rip currents by setting the offshore length threshold of the sediment plume segment, but this approach can only apply to rip currents with sediment plumes.

When the detection advantages of deep convolutional neural networks in the marine field are manifested, the use of CNN for rip current detection has also become a hot issue. In [28], Gregory Perrier invented a patented device for automatic rip currents detection, which uses the collected various images with rips under the help of neural network to imitate human perception to identify rip currents. This method has a certain detection effect on rip currents that leave the coast vertically, but ignores the various forms of rip currents, as there are some rip currents not necessarily perpendicular to the coast.

Time-averaged image is also applied to the detection of rip currents, in [29], Corey Maryan et al explored state-of-the-art machine learning algorithms at the time, including the Viola-Jones algorithm, convolutional neural networks, and classifiers on rip current image data sets. And apply machine learning to identify rip currents in time-averaged images. Besides, they found five new Haar features of rip currents, which successfully supplemented the original Haar feature sets and provided valuable data for the detection of rip currents.

In [21], De Silva et al proposed a more feasible detection scheme based on convolutional network, and for the first time created a public data set on rip currents, using a two-stage Faster-RCNN network and a custom frame aggregation stage for detection from still images or videos. Its measurement accuracy is higher than other rip currents detection methods. However, it is only an application based on convolutional networks without improvement in the network structure. And the detection speed of the two-stage model Faster-RCNN is still much slower than the one-stage model YOLO. Due to the model is too large, it cannot meet the requirements of rapidity and lightness of real-time detection.

3 YOLOv5s Model

There are four versions of YOLOv5 in which the network depth and width increase in turn. Among the YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x versions, YOLOv5s has the smallest weight file, only 14MB, which is in line with the characteristics of lightweight networks. Compared with other YOLO series networks, the average detection accuracy is significantly reduced, but the training time is greatly shortened, which can meet the basic requirements of real-time detection, and realize the timely and early warning of rip currents. Therefore, we choose YOLOv5s as the baseline model. It is structurally divided into four parts: Input End, Backbone, Neck, and Detection Head. The specific structure is shown in Fig. 1.

YOLOv5s connects a Focus module to the input end, expands the input channel by 4 times through the slicing operation, then obtains a double down-sampling feature map without information loss through the convolution operation, which can effectively improve the speed while reducing the amount of calculation. In feature extraction, the CBL (Conv+BN+Leaky_Relu) module is used for convolution, normalization, and activation, and the C3 (CBL+Bottlneck+Concat) module strengthens extraction and optimizes gradients to speed up network reasoning. Finally, the SPP module will unify the network output size and reduce the impact of image resize when the input image size is inconsistent.

YOLOv5s adopts the method of FPN (Feature Pyramid Network) combined with PAN (Pant Aggregation Network) [30]. In the Neck area, FPN upsamples the image from top to bottom, and combines the extracted fea-

tures with the features extracted by the backbone to enhance feature information. In order to better fuse semantic features with location information, a bottom-up PAN feature pyramid structure is constructed behind the FPN for passing the underlying location information upwards. Through the combined network structure of FPN and PAN, the network model obtains more abundant feature information.

Finally, the image is divided into three grid sizes of 20x20, 40x40, and 80x80, which are used to detect large, medium, and small objects respectively. Each grid has a corresponding vector, which contains the prediction type and category confidence respectively.

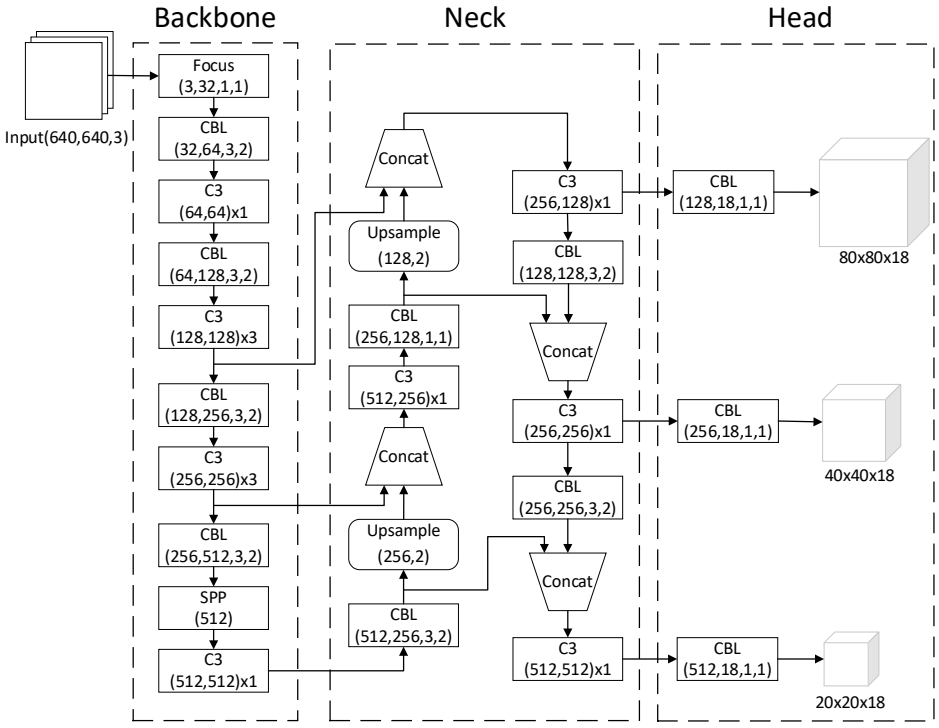


Fig. 1. YOLOv5s network structure

4 Improvement Methods

4.1 Joint Dilation Convolution Module

The prediction of YOLOv5s is based on the feature map of the last layer. The number of pixels that a point can map to the original image determines the maximum size of the target that the model can detect. Since the size of the rip current targets are large in the images, and extracting more rip current features is very important for the reasoning of the network, the receptive field of the model needs to be expanded. There are two ways to expand: a) Image downsampling. But it will lose some feature informations, which reduces the useful informations in disguise; b) Increase the number of convolutional layers. The superposition of convolutional layers can greatly expand the receptive field, but as the number of layers increases, the amount of network calculation is also greatly increased. Consequently, it will inhibit the calculation speed of the model.

In order to effectively solve the problem of expanding the receptive field. The literature [31] proposed the dilated convolution. Dilated convolution adds holes on the basis of ordinary convolution, that is, complements each two parameters of the basic convolution with 0. The dilation rate parameter is used to control the number of interval 0s. The calculation formula of the size of the convolution kernel after expansion is expressed as:

$$f = d(k-1)+1. \tag{1}$$

In the formula, *f* is the size of the convolution kernel after expansion, is the dilation rate, and *k* is the size of the original convolution kernel. When dilation rate = 2, the 3x3 convolution kernel is expanded into a 5x5 convolution kernel, which effectively expands the receptive field while keeping the parameters unchanged. As shown in Fig. 2.

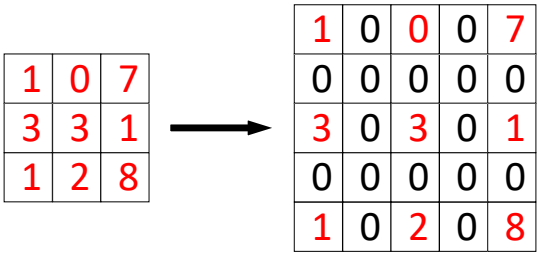


Fig. 2. Dilation convolution of division rate = 2

Dilated convolution kernel stacking can expand the receptive field without losing the size of the feature map. However, when the superimposed convolution kernels all have the same expansion rate, due to the gap between the parameters, there will be a discontinuity based on the convolution center on the overall feature map, and a "checkerboard" effect will occur, appearing outward from the convolution center. The expanded grid state, as shown in Fig. 3, is the superposition result of the 3x3 convolution kernel with three layers of dilation = 2. The color depth indicates the number of feature extractions.

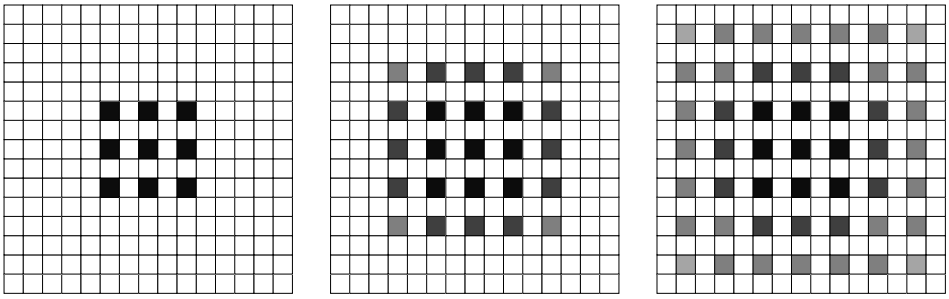


Fig. 3. "Checkerboard" effect

In view of the characteristics and problems of dilated convolution, we design a JDC (Joint Dilated Convolution) module. JDC uses three convolution kernels with sizes of 1x1, 3x3, and 5x5 to perform convolution operations in three groups of channels. The specific module structure is as follows shown in Fig. 4. The 1x1 convolution kernel in the structure is to implement dimensionality reduction to reduce the amount of parameters, and the 3x3 and 5x5 convolution kernels are expanded convolution superposition for feature extraction operations. In order to avoid the "checkerboard" effect caused by incomplete input feature extraction, the superimposed dilated convolution kernels should not have the same dilation rate, nor can they be superimposed in multiples, so the dilation rate of each layer is set to an equal difference column to make up for the omission of feature information when superimposed by dilated convolution. The expansion rate of the two-layer dilated convolutional layers of 3x3 and 5x5 is uniformly set to d(dilation) = [2, 3], p(padding) is set to [2, 3], [4, 6] respectively, and the last three groups of channels are spliced and output.

We embed the JDC module into the horizontal connection between the Backbone area of YOLOv5s and the FPN feature pyramid in the Neck area, reducing the amount of parameters as much as possible while fully expanding the receptive field and increasing the adaptability of the network to scale, which can effectively and efficiently extract features, and fusion to improve the accuracy of detection.

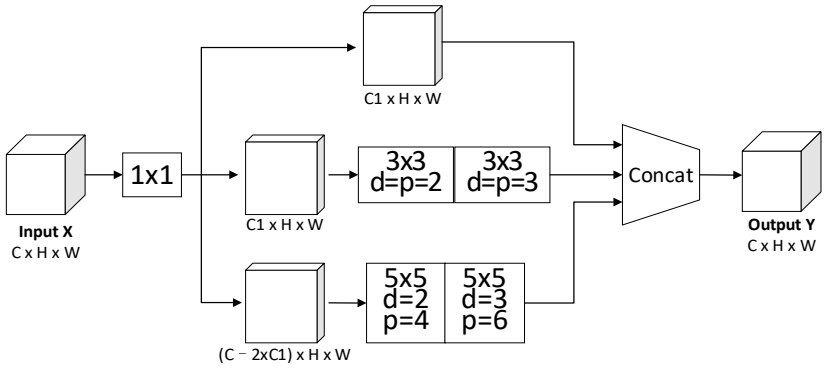


Fig. 4. JDC module structure

4.2 Parameter-Free Attention Module

In order to let the network focus more on filtering out the high-value rip current features information from the input image and suppress other useless information, we introduce an attention mechanism to multiply the effective features of the rip currents with the weights to improve its importance, and enhance the processing of feature information by the network. CBAM (Convolutional Block Attention Module) [32] infers the attention weights in turn along the two dimensions of space and channel, and then multiplies with the original feature map to adaptively adjust the features. CoordAttention (Coordinate Attention) [33] decomposes the channel attention into two parallel 1D feature coding processes, instead of 2D global pooling to convert the feature tensor into a single feature vector, avoiding the introduction of positional information loss. The above attention mechanisms have achieved good results in convolutional neural networks, but due to complex operations such as pooling, the structure is relatively bloated, and it is mainly aimed at improving the recognition accuracy of small targets. In order to keep the model simple enough to be embedded in the device, on the basis of effectively enhancing the large target feature, and reducing the introduction of module parameters to balance the calculation cost, we introduces a parameter-free lightweight attention mechanism SimAM after the JDC module.

Based on neuroscience, the energy function for each neuron is defined as:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x}_i)^2 + (y_t - \hat{t})^2 \ . \tag{2}$$

Among them, $\hat{t} = w_t t + b_t$ and $\hat{x}_i = w_t x_i + b_t$ are the linear transformations of target neuron t and other neurons x_i in the same channel of the input feature, w_t and b_t are the weights and biases of the linear transformation, i is the index in the spatial dimension, and y_t and y_0 are the two different values. $M = H \times W$ is the number of neurons on a channel. Minimizing Eq. (2) is to find the linear separability of t and x_i in the same channel. The smaller the final added value, the greater the difference between the two, and the smaller the overlap, the more important it is to represent target neuron t. Replacing y_t and y_0 with binary labels (1 and -1) and adding regularization can get the final energy function:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2 \ . \tag{3}$$

Solving the above formula can get the solutions of w_t and b_t:

$$w_t = \frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \quad . \tag{4}$$

$$b_t = -\frac{1}{2}(t+\mu_t)w_t \quad . \tag{5}$$

μ_t is the mean and σ^2 is the variance:

$$\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \quad . \tag{6}$$

$$\sigma^2 = \frac{1}{M-1}\sum_{i}^{M-1}(x_i - \mu_t)^2 \quad . \tag{7}$$

Substituting Eq. (4) and Eq. (5) into Eq. (3) can obtain the formula for the minimum value of the energy function:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\sigma^2 + 2\lambda} \quad . \tag{8}$$

According to the inference of Eq. (8), the lower the energy value, the more important neuron t will be, which is inversely proportional. Therefore, $\frac{1}{e_t^*}$ can be used to describe the importance of neurons. According to the definition of the attention mechanism, $\frac{1}{e_t^*}$ is constrained by the Sigmoid function, and finally enhances image features as weights:

$$\tilde{X} = Sigmoid(\frac{1}{E}) \otimes X \quad . \tag{9}$$

After the SimAM module is connected to the JDC module to extract the features of rip currents, the effective features are strengthened, so the network can focus on processing the effective features, which can improve the efficiency and accuracy of the detection. Meanwhile, the parameter-free feature of SimAM avoids the introduction of too much computation, keeps the model simple, and does not affect the detection rate of the model.

4.3 Network Structure Adjustment

In view of the fact that the area of the rip currents area occupies a large proportion in the image, the number of AnchorBox can be appropriately reduced to abandon the detection of small targets, and the model complexity can be reduced by deleting the 80x80 feature map branch of the Neck area of the YOLOv5s model to improve the detection speed. Since the Mosaic method appears redundant in the structure after removing the 80×80 feature map branch, it is not enabled in the improved model. Finally, the JDC and SimAM modules are connected and embedded into the FPN lateral connection of YOLOv5s. The model structure is shown in Fig. 5:

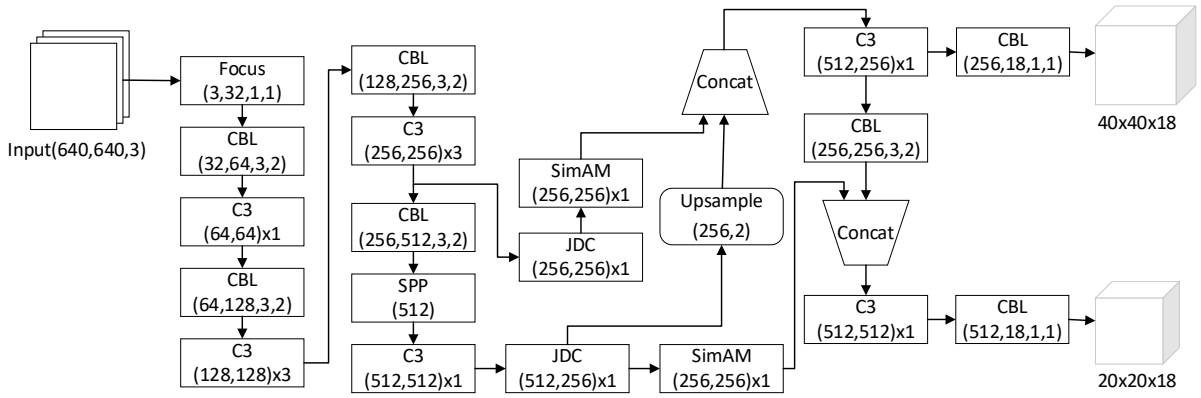


Fig. 5. Improved YOLOv5s network structure

5 Experiment and Analysis

5.1 Experimental Environment

The environmental hardware experiment: CPU is Intel(R) Core(TM) i5-10400F, 16GB memory; GPU is NVIDIA GeForce RTX 3060Ti, 8GB video memory. Software environment: Windows10 operating system; Python version 3.9.0; deep learning framework using Pytorch 1.9.0; IDE is Pycharm.

5.2 Data set and Parameters

The rip currents data set disclosed by Akila et al [21] is used to test the effect of the improved model. It contains 1,780 images. These images are classified by professional rips researchers, of which 1,628 images are with rip currents and 152 images do not. A partial sample of this data set is shown in Fig. 6.

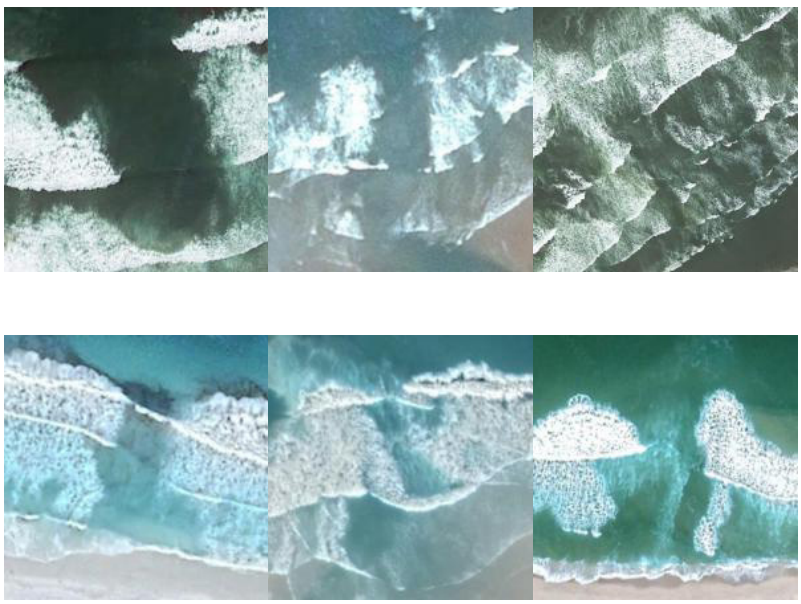


Fig. 6. Schematic diagram of some samples of data sets

The data set is divided into training set, validation set and test set according to the ratio of 8:1:1 as 1424, 178 and 178 respectively. The parameters shown in Table 1 are set respectively in several mainstreams. The data set is trained on the target detection model, and experimental results are compared and analyzed with the improved model. More details will be presented in Section 5.4.

Table 1. Experimental parameters

Parameters	Value
Image_size	640x640
Channel	3
moment	0.937
Ir	0.01
Weight_decay	0.0005
Batch_size	32
epoch	300

5.3 Evaluation Indicators

We use the following four evaluation indicators to verify the performance and detection effect of the improved YOLOv5s model: P (Precision), R (Recall), mAP (mean Average Precision), FPS (Frames Per Second).

$$\mathrm{Pr}\,cision = \frac{TP}{TP + FP} \quad . \tag{10}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \quad . \tag{11}$$

$$mAP = \frac{\sum \int_0^1 P(R)dR}{Nc} \quad . \tag{12}$$

TP is true positive, indicating the number of correctly detected targets. FP is false positive, indicating the number of wrongly detected targets, and FN is false negative, indicating that positive targets are incorrectly predicted as negative. Nc represents the number of target categories. Generally, the higher the mAP is, the better the model detection will effect. FPS is used to measure the detection speed of the model. The larger the FPS is, the faster the model can process images.

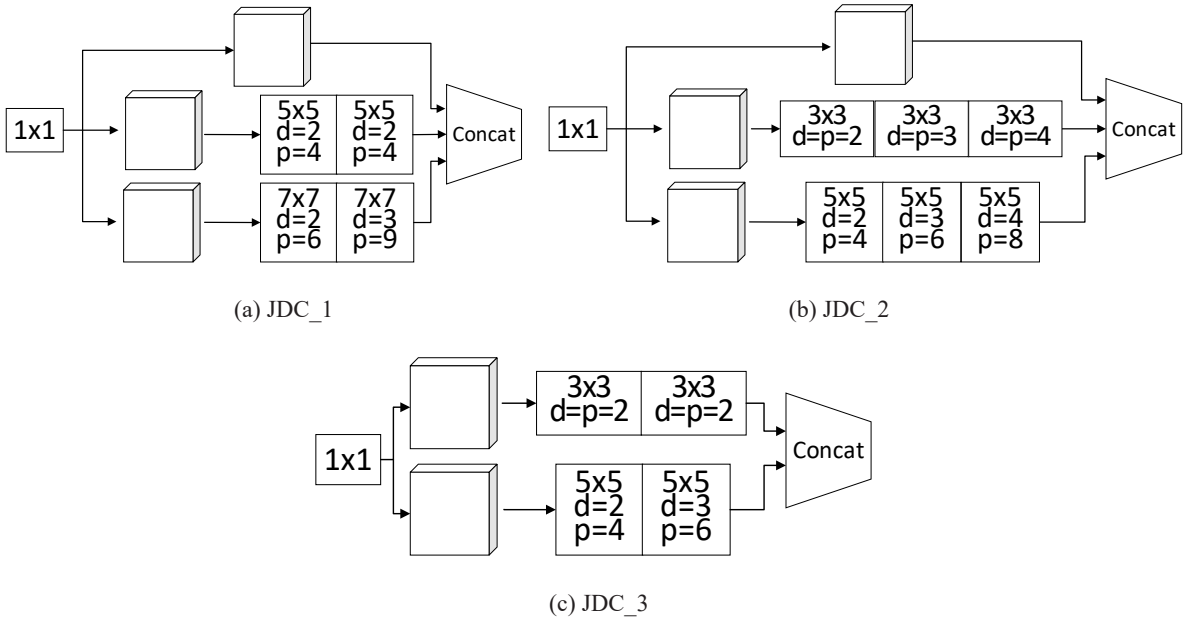
5.4 Results and Analysis

Detection Scale Comparison. The original YOLOv5s network has three detection scales of 20x20, 40x40, and 80x80. Based on the characteristic of rip currents being a large target in detection, we deletes the 80x80 feature map detection scale, and retains the detection scale of medium and large target. Table 2 shows the training results of the network model before and after the change. After deleting the small target detection scale branch, the model is reduced by 0.54MB, and the mAP is decreased by 0.33%, but the detection speed can reach up to 49.63 frames per second, which is 3.58 higher than the original model. Therefore, we retain the two feature map detection scales of 20x20 and 40x40, and further improve the original model on this basis.

Table 2. Comparison of detection scale experiments

Network model	Model size/ MB	mAP@0.5/ %	FPS/ frames · s^{-1}
Original scale	6.74	88.15	46.05
Modified scale	6.20	87.82	49.63

Different JDC Structure Comparison. We design three other JDCs with different structures to compare the detection performance of the models under different combinations. On the basis of JDC, JDC_1 removes the branches that are not processed by dilated convolution, and explores the necessity of performing full-channel dilated convolution; JDC_2 keeps the dilation rate unchanged, and increases the convolution kernel to 5x5 and 7x7; JDC_3 adds a layer convolution kernel. The latter two structures use different methods to further expand the receptive field and explore the best receptive field range suitable for images with rips. Three joint dilated convolution structures are shown in Fig. 7.



(a) JDC_1

(b) JDC_2

(c) JDC_3

Fig. 7. The other three joint extended convolution structures

Different joint dilated convolution structures are applied to the network model with the 80x80 feature map branch removed for testing. The experimental results are shown in Table 3. The results show that several JDC structures proposed above can increase the mAP and FPS, and can improve the model performance. Among the three structures, JDC_1, which performs full-channel joint dilated convolution, improves the mAP and FPS the most, but the model size also increases the most, reaching 1.64MB. JDC_2 and JDC_3 use two ways of expanding the convolution kernel and stacking the convolution layer to expand the receptive field, but the improvement effect of the model is not obvious, indicating that the receptive field exceeds the optimal expansion range suitable for rip current targets, and it is relatively increased the number of model parameters greatly. Compared with the above structure, original JDC increases the mAP of the model by 2.99%, reaching 91.14%, while the FPS increases by 2.05 frames per second, and the model size increases by only 0.45MB, which is the best among the above structures. So we choose the original JDC to embed the Neck area.

Table 3. Training results of different joint expansion convolution structures

Network model	Model size/ MB	mAP@0.5/ %	FPS/ frames · s^{-1}
Original	6.74	88.15	46.05
JDC_1	8.38	90.29	47.54
JDC_2	8.30	88.41	46.13
JDC_3	7.66	89.48	46.68
JDC	7.19	91.14	48.10

Attention Mechanism Comparison. Based on the improved model in the first two sections, we conduct a comparative experiment by integrating the attention mechanism CoordAttention and CBAM after the JDC module to verify the compatibility of SimAM and the model. The experimental results are shown in Table 4. Although the CBAM module can capture the local correlation of feature information, it is difficult to capture the dependence on large regions. The convolution operation in CoordAttention further expands the receptive field, embeds location information into channel attention for small targets, and generates coordinate attention to enhance feature aggregation. Adding CoordAttention and CBAM modules can improve the model's detection accuracy for small targets, but neither of them is suitable for large targets such as rip currents. The introduction of additional parameters to increase the number of network layers or the convolution operation exceeds the optimal receptive field. It will be counterproductive, as the detection is becoming more accuracy and the FPS is decreasing. Unlike the combination of spatial attention and channel attention, SimAM explores the importance of each neuron to generate attention weights. Compared with the above two modules, there is no convolution operation and no additional parameters. The introduction can effectively target the rip currents target. As shown in the table, after adding the SimAM module, the mAP is increased by 1.01%, and the FPS is increased by 0.13 frames per second.

Table 4. Training results of different joint expansion convolution structures

Network model	Model size/ MB	mAP@0.5/ %	FPS/ frames · s^{-1}
Orginal	6.74	88.15	46.05
JDC_1	8.38	90.29	47.54
JDC_2	8.30	88.41	46.13
JDC_3	7.66	89.48	46.68
JDC	7.19	91.14	48.10

Ablation Experiment. The improvement scheme in Section 4 is subjected to ablation experiments to verify the impact of the improvement of the modules on the performance of the model. The experimental results are shown in Table 5. It can be seen that in experiment 3, after adding the JDC module to the original network, the accuracy increased by 1.65%, and the detection rate increased by 0.63 frames per second, indicating that JDC can improve the feature extraction ability of the model. But the FPS is still the same without the scale improvement or the improvement is not obvious enough. Compared with experiment 3, by adding scale improvement in experiment 5, the model is reduced by 0.55MB, the detection rate is increased by 1.62 frames per second, and the mAP is also increased by 1.34%, indicating that removing redundant network detection branches can significantly improve network detection capabilities. In experiment 8, the parameter-free attention mechanism SimAM was added to further enhance the feature processing capability of the network. Compared with experiment 5, the model size remained unchanged, but the mAP has increased by 1.01%, and the detection rate has increased by 0.13 frames per second.

Table 5. Ablation experimental results

Numbers	Scale improvement	JDC	SimAM	Model size/ MB	mAP@0.5/ %	FPS/ frames · s^{-1}
1				6.74	88.15	46.05
2	√			6.19	87.82	49.63
3		√		7.74	89.80	46.48
4			√	6.74	88.89	43.13
5	√	√		7.19	91.14	48.10
6	√		√	6.19	88.21	47.14
7		√	√	7.74	90.13	42.08
8	√	√	√	7.19	92.15	48.23

To sum up, compared with the original model, the improved YOLOv5s model only increases the size by 0.45MB, the mAP increases by 4%, and the FPS also increases by 2.18 frames per second. The mAP comparison curve shown in Fig. 8 indicate our model is overall higher than the original model, which shows our improvement scheme can effectively improve the detection accuracy of the model.

When detecting videos containing rip currents, the comparison between the detection effect of the original YOLOv5s model and the improved model is shown in Fig. 9. It can be seen from the figure that due to the insufficient extraction and processing capabilities of the original model for rip current features before the improvement, there will be false detections, missed detections, and overlapping marks in the same area during the rip currents detection. The improved model can avoid the above problems and achieve better recognition accuracy. Based on the above experimental verification, the improved model is more suitable for the detection and identification of rip currents than the original model.

In addition to the above video-based detection, this paper also uses drones to shoot some static images on the coast to test the accuracy of the recognition. The results are shown in Fig. 10. The red frame represents the real frame, and the yellow frame represents the prediction frame. An image without a bounding box represents no rip currents.

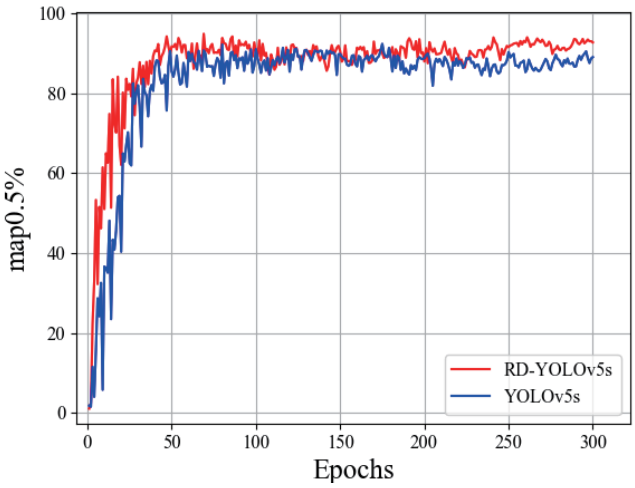


Fig. 8. Comparison of map before and after improvement

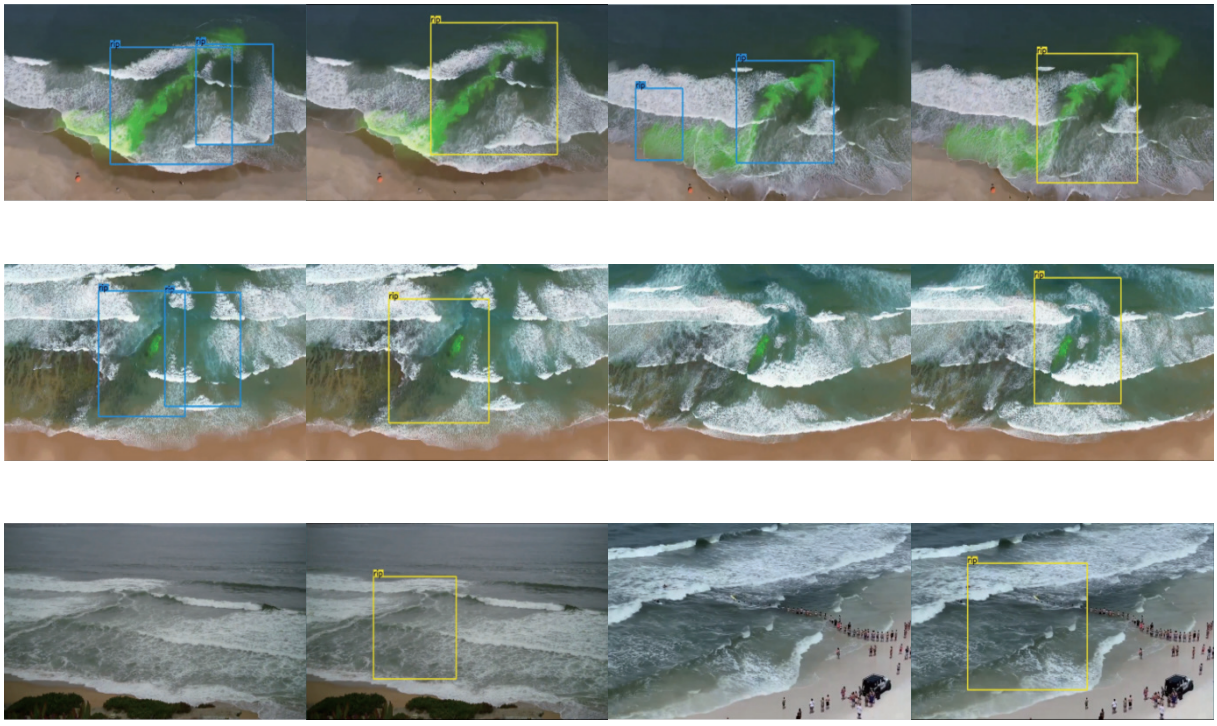


Fig. 9. Actual test comparison of some frames in the video of the original model (left) and the improved model (right)

Fig. 10. Detection based on aerial images

Experimental Comparison on Different Models. Under the same experimental environment and data set, using the same data division strategy and parameter settings, the improved model in this paper is the same as YOLOv5s, YOLOv5m, YOLOv5l, YOLOv4-tiny, YOLOv4, YOLOv3, YOLOv3-tiny, YOLOX-s, and Faster R-CNN. It is trained separately for 300 iterations together with existing object detection algorithms such as Efficiendet. The related results are shown in Table 6.

Table 6. Ablation experimental results

Network model	mAP@0.5/ %	FPS/ frames · s^{-1}
YOLOv5s	88.16	46.05
YOLOv5m	85.84	41.68
YOLOv5l	84.89	38.31
YOLOv4	91.69	45.07
YOLOv4-tiny	67.22	47.65
YOLOv3	84.93	46.95
YOLOv3-tiny	75.68	49.21
YOLOX-s	86.45	55.76
Faster R-CNN	48.96	16.18
Efficientdet-d0	55.14	23.26
Efficientdet-d1	86.86	18.55
Efficientdet-d2	87.81	17.33
Ours	**92.15**	**48.23**

The experimental results show the improved YOLOv5s model has a certain improvement in the average detection accuracy on mAP and FPS compared with other mainstream network models. Comparing YOLOv3-tiny and YOLOX-s, although the FPS is reduced by 0.98 frames per second and 7.53 frames per second, the mAP is increased by 16.47% and 5.7% respectively. Especially for the two-stage model Faster R-CNN, the mAP and FPS value have been greatly improved, which once again verifies the feasibility of the improved model, reflecting its performance in rip currents. It has more advantages than other mainstream models in target detection.

6 Conclusion

By designing the joint expansion convolution JDC module, this research solves the problem of large increase in parameters or loss of feature information caused by the expansion of receptive field in the past, and effectively expands the receptive field to extract the feature information of rip current targets and carry out multi-scale fusion. Then, the attention mechanism SimAM is added without introducing additional parameters to keep the model concise and strengthen effective feature processing to further improve the detection accuracy. Finally, the 80x80 detection branch is removed, which reduces the complexity of the network, speeds up the detection speed, and makes the model adapt to the detection of large targets such as rip currents. Through the above improvements, the experimental test results on the rip current data set disclosed by Akila show that the average detection accuracy of the model reaches 92.15%, the detection rate reaches 48.23 frames per second, and the size of the model is only 7.20 MB, which proves our model has a good improvement effect, allows the application of embedded devices and meets the requirements of real-time detection.

At present, we only completed the detection of a single target. In the future, the data set will be expanded, training samples will be added, and the multi-target detection method will be improved, that is, multiple rip current targets in a image will be detected. Secondly, the research on target detection of rip currents based on deep learning is in a lack period, and other features such as the category and speed of rip currents are still difficult to detect in a convenient way, including the loss of target details due to the influence of external factors such as strong sea breeze and easy exposure of images when obtaining the images with rips. In view of the above problems, it is necessary to preprocess and reconstruct the images. Whether we can further detect and identify the rip currents under the complex background will become the key research direction of the follow-up work.

7 Acknowledgement

This research was supported by the National Science Foundation of China No. 42176167.

References

[1] F.P. Shepard, K.O. Emery, E.C.L. Fond, Rip Currents: A Process of Geological Importance, The Journal of Geology 49(4)(1941) 337-369.
[2] A.J. Bowen, Rip currents: 1. theoretical investigations, Journal of Geophysical Research 74(23)(1969) 5467-5478.
[3] A.D. SHORT, Australian rip systems—friend or foe? Journal of Coastal Research (50)(2007) 7-11.
[4] C. Brannstrom, H.L. Brown, C. Houser, S. Trimble, A. Lavoie, "You can't see them from sitting here": Evaluating beach user understanding of a rip current warning sign, Applied Geography 56(2015) 61-70.
[5] J.A. Brown, J.H. MacMahan, A.J.H.M. Reniers, E.B. Thornton, Field observations of surf zone–inner shelf ex-change on a rip-channeled beach, Journal of Physical Oceanography 45(9)(2015) 2339-2355.
[6] D. Buscombe, R.J. Carini, A data-driven approach to classifying wave breaking in infrared imagery, Remote Sensing 11(7)(2019) 859.
[7] J.Y. Luo, J.-O. Irisson, B. Graham, C. Guigand, A. Sarafraz, C. Mader, R.K. Cowen, Automated plankton image analysis using convolutional neural networks, Limnology and Oceanography: Methods 16(12)(2018) 814-827.
[8] A.G. Howard, M.L. Zhu, B. Chen, D. Kalenichenko, W.J. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: efficient convolutional neural networks for mobile vision applications. <https://arxiv.org/abs/1704.04861>, 2017 (accessed 13.10.21).
[9] M.X. Tan, R.M. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, in: Proc. of the 2020 Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[10] F.-N. Iandola, S. Han, M.-W. Moskewicz, K. Ashraf, W.-J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. <https://arxiv.org/abs/1602.07360>, 2016 (accessed 07.08.21).
[11] S. Mehta, M. Rastegari, MobileVit: light-weight, general-purpose, and mobile-friendly vision transformer. <https://arxiv.org/abs/2110.02178v2>, 2016 (accessed 23.09.21).
[12] B. Yan, P. Fan, X. Lei, Z. Liu, F. Yang, A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5, Remote Sensing 13(9)(2021) 1619.
[13] C. Guo, X.-L. Lv, Y. Zhang, M.-L. Zhang, Improved YOLOv4-tiny network for real-time electronic component detection, Scientific Reports 11(2021) 22744.
[14] W. Zhan, C. Sun, M. Wang, J.H. She, Y.Y. Zhang, Z.L. Zhang, Y. Sun, An improved YOLOv5 real-time detection method for small objects captured by UAV, Soft Computing 26(2022) 361-373.

[15] Y.K. Li, X. Lv, P.P. Huang, W. Xu, W. Tan, Y.F. Dong, SAR Ship Target Detection Based on Improved YOLOv5s, in: Proc. of the 10th International Conference on Control, Automation and Information Sciences (ICCAIS), 2021.
[16] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[17] R. Girshick, Fast R-CNN, in: Proc. of the IEEE International Conference on Computer Vision (ICCV), 2015.
[18] S.Q. Ren, K.M. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6)(2017) 1137-1149.
[19] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[20] L.X. Yang, R.-Y. Zhang, L.D. Li, X.H. Xie, SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks, in: Proc. of the 38th International Conference on Machine Learning (PMLR), 2021.
[21] A.D. Silva, I. Mori, G. Dusek, J. Davis, A. Pang, Automated rip current Detection with Region based Convolutional Neural Networks, Coastal Engineering 166(2021) 103859.
[22] Surf Life Saving Australia, National coastal safety report 2019. <https://www.surflifesaving.com.au/wp-content/uploads/sites/2/2021/07/SLSA-National-Coastal-Safety-Report-2019-2.pdf>, 2019 (accessed 07.05.21).
[23] D.B. Clark, L. Lenain, F. Feddersen, E. Boss, R. Guza, Aerial imaging of fluorescent dye in the near shore, Journal of Atmospheric and Oceanic Technology 31(6)(2014) 1410-1421.
[24] S.B. Leatherman, S.P. Leatherman, Techniques for detecting and measuring rip currents, International Journal of Earth Science and Geophysics 3(1)(2017) 1-5.
[25] B. Castelle, R. Almar, M. Dorel, J.-P. Lefebvre, N. Senechal, E.J. Anthony, R. Laibi, R. Chuchla, Y.D. Penhoat, Rip currents and circulation on a high-energy low-tide-terraced beach (Grand Popo, Benin, West Africa), Journal of Coastal Research 70(sp1)(2014) 633-638.
[26] S.J. Philip, A. Pang, Detecting and Visualizing rip current Using Optical Flow, in: Proc. Eurographics Conference on Visualization (Euro Vis), 2016.
[27] Y.L. Liu, C.H. Wu, Lifeguarding Operational Camera Kiosk System (LOCKS) for flash rip warning: Development and application, Coastal Engineering 152(2019) 103537.
[28] G. Perrier, Automated rip tide detection system, US Patent No. 6,931,144, 2005.
[29] C. Maryan, M.T. Hoque, C. Michael, E. Ioup, M. Abdelguerfi, Machine learning applications in detecting rip channels from images, Applied Soft Computing Journal 78(2019) 84-93.
[30] S. Liu, L. Qi, H.F. Qin, J.P. Shi, J.Y. Jia, Path aggregation network for instance segmentation, in: Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[31] F. Yu, V. Koltun, Multi-Scale Context Aggregation by Dilated Convolutions, in: Proc. International Conference on Learning Representations (ICLR), 2016.
[32] S.H. Woo, J.C. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proc. the European Conference on Computer Vision (ECCV), 2018.
[33] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.