Hindawi
Mobile Information Systems
Volume 2021, Article ID 3168464, 10 pages
https://doi.org/10.1155/2021/3168464

Research Article

A Shadow Capture Deep Neural Network for Underwater Forward-Looking Sonar Image Detection

Taowen Xiao,[1] Zijian Cai,[1] Cong Lin,[1] and Qiong Chen[2]

[1]*College of Electronics and Information Engineering, Guangdong Ocean University, Zhangjiang 524025, China*
[2]*Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Institute for Global Change Studies, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Cong Lin; lincong@hainanu.edu.cn and Qiong Chen; qiongchen@mail.tsinghua.edu.cn

Received 25 October 2021; Accepted 7 December 2021; Published 30 December 2021

Academic Editor: Han Wang

Image sonar is a widely used wireless communication technology for detecting underwater objects, but the detection process often leads to increased difficulty in object identification due to the lack of equipment resolution. In view of the remarkable results achieved by artificial intelligence techniques in the field of underwater wireless communication research, we propose an object detection method based on convolutional neural network (CNN) and shadow information capture to improve the object recognition and localization effect of underwater sonar images by making full use of the shadow information of the object. We design a Shadow Capture Module (SCM) that can capture the shadow information in the feature map and utilize them. SCM is compatible with CNN models that have a small increase in parameters and a certain degree of portability, and it can effectively alleviate the recognition difficulties caused by the lack of device resolution through referencing shadow features. Through extensive experiments on the underwater sonar data set provided by Pengcheng Lab, the proposed method can effectively improve the feature representation of the CNN model and enhance the difference between class and class features. Under the main evaluation standard of PASCAL VOC 2012, the proposed method improved from an average accuracy (mAP) of 69.61% to 75.73% at an IOU threshold of 0.7, which exceeds many existing conventional deep learning models, while the lightweight design of our proposed module is more helpful for the implementation of artificial intelligence technology in the field of underwater wireless communication.

1. Introduction

The electromagnetic wave, light wave, and sound wave are commonly used in the world as the wireless communication carrier. In the field of wireless communication, seawater contains a variety of minerals and becomes a conductive medium. The electromagnetic wave propagating to the ocean will be blocked by seawater, resulting in the rapid waste of most energy. Therefore, the electromagnetic wave propagating in seawater will be greatly suppressed. Based on the above analysis, the electromagnetic wave transmitting in the seawater will be greatly refrained, because most of its energies will be wasted quickly and spreading to ocean will be blocked by the sea water. Light is also an electromagnetic wave in essence, so its transmission in seawater will also be limited [1–3]. Since the reason that the sound wave is a mechanical wave that can travel in the elastic medium of seawater without much interference, it is widely used in the field of underwater wireless communication. Inspired by sound waves, many sonar devices have been developed to measure depth of water or detect underwater objects. For example, forward-looking sonar equipment can obtain the reflection information of sound waves and generate high-resolution sonar images, which is usually used to obtain underwater information in the form of images.

Sonar image object recognition methods are mainly divided into traditional mathematical modelling methods and detection methods based on convolutional neural networks (CNN). Traditional mathematical modelling methods can handle more sonar object recognition tasks,

and it uses image process methods such as Scale Invariant Feature Transform (SIFT), Directional Gradient Histogram (HOG) [4], and Fisher vectors to extract object features and then pass machine learning or pattern matching classifies object features [5–7]. The features extracted by this method can perform well for specific data sets and tasks, but the generalization ability of most features is limited, and feature extraction still needs professional knowledge and a lot of experiments. In contrast, the convolutional neural network (CNN) optimizes its parameters with back gradient propagation, which makes it possible to combine feature extraction and model prediction into the same pipeline.

As a powerful image classification and object detection model, the detection method based on CNN is one of the most popular deep learning structures [8–12]. Compared with traditional computer vision methods, the CNN method is more widely used in the field of sonar image recognition since the advantages of automatic feature extraction and multilevel feature extraction [13–20]. There are two types of application of CNN model. One is to combine the convolution module or the fully connected layer module to build a sonar image recognition model. Matias [13] and Valdenegro-Toro [14] both designed a CNN model for object recognition in sonar images, which achieved better results than pattern matching methods. Williams [15] designed a deep convolutional neural network and applied it to multiple binary classification tasks to distinguish different categories in sonar data sets, which is better than the original classifier based on manual features. The other type of CNN model is to use advanced theory and experience in computer vision (CV) to modify some excellent models from CV to meet the needs of sonar image recognition. Galusha et al. [16] use deep convolution neural network for object detection such as object recognition and location, and the pixels of SAS image are reduced to RoIs (region of interests). This detection form resembles the standard detection stream of the two-stage model of object detection. Breistein [17] and Neves et al. [18] introduced a one-stage object detection model yolov2 for sonar image recognition, and the model achieved expected results in their respective data sets. Zacchini et al. [19] introduced Mask RCNN for sonar image recognition and object localization and tested the functions of the model at LaSpezia Naval Support and Laboratory Center. Fan et al. [20] reduced network parameters by modifying the network structure of the model without affecting accuracy.

The above CNN methods have different emphasis on using deep learning models. Although good results have been achieved in their respective data sets, they can only locate and recognize one sonar object due to the small number of data sets in most cases. How to effectively increase the feature gap between different categories and narrow the feature gap between the same category is particularly important. In addition, we noticed that the sound waves emitted by the forward-looking sonar device will bounce off the objects it touches and cannot reach the area behind the object, thus forming a shadow related to the object shape on the generated sonar image. It is difficult for traditional image processing methods to interpret the highlight area representing the object due to the above-mentioned imaging defects of sonar images, so shadow information is often introduced as an additional auxiliary identification feature [21–24]. Although shadow feature has been given considerable attention in previous recognition tasks, the methods proposed by them are exclusive, which means that it is impossible to separate the extraction steps of shadow features and transplant them to the deep learning method. Therefore, researchers have not paid enough attention to the shadow information of sonar images in the research of existing deep learning methods.

In order to make full use of the shadow information in the forward-looking sonar image and improve the accuracy of image object recognition and positioning, a neural network detection method based on shadow capture module (SCM) is designed. SCM module can capture the shadow features in the image according to the characteristic that the shadow appears directly above the object, which is very lightweight and portable. In addition, we made a data set with shadow as the object to train and test the ability of the original model to distinguish shadows. Through a large number of experiments on underwater sonar images, it is found that the detection accuracy of the network model added with SCM module is 6.12% higher than the original model, which exceeds the existing common CNN model. The main contributions of this paper can be summarized as follows:

(1) In order to utilize the shadow information in the forward-looking sonar image, we propose a structure based on CNN model that can capture the shadow information and integrate the shadow information into the feature map.

(2) We made a data set with the object shadow to train the proposed model and then used them to identify and locate the shadow of the object. The final result showed that the detection accuracy reached 94%, which proved that the CNN model can also adapt to sonar image shadow recognition and provides a priori basis for our method.

(3) The proposed module SCM in this paper can effectively increase the difference degree of features between categories of deep learning model in multicategory detection tasks. After the addition of the designed module, the accuracy of CNN model is improved by 6.12%, which is higher than the mainstream object detection model in CV field.

2. Shadow Capture Network

There will be a shadow associated with its shape directly above the target object in the forward-looking sonar picture. It is an undetected area formed by the object blocking the sound waves emitted by the sonar device. The shape of this area varies with the shape of the object. Although the angle at which the sonar device emits sound waves will affect the shape of the shadow to a certain extent, the only thing that plays a dominant role in the shape of the shadow is the shape of the object when the emission angle is basically constant.

Therefore, in the sonar picture generated by a forward-looking sonar device emitting sound waves at a stable angle, the shadow can be regarded as another characteristic expression of the object.

In this section, we describe the rationale for proposing this structure and the details of its construction. As with CenterNet [25], our model eventually outputs a heat map that contains both the classification prediction score of the object and the coordinates of the object's position in the heat map. Since there is a one-to-one mapping relationship between the object points in the heat map and the original image, the position of the object points in the heat map relative to the heat map represents the position of the object in the original image relative to the original image. The use of shadow features in the forward-looking sonar image can serve to increase the feature differences between categories and improve the response values of object points in the heat map, thus improving the network's ability to recognize objects. Since the classification scores of hotspots are generated simultaneously with the hotspot coordinates, a higher hotspot response value also means that the network is more certain about the location of the object, so the optimization of the classification effect also represents an enhancement of the localization effect. Therefore, capturing shaded features and fusing them into the feature maps in the network can fully improve the prediction of heat maps by our model.

Therefore, we replace the module in CenterNet that is used to predict the object with a feature capture module that captures the shadow features in the feature map, and also fuse the captured shadow features into the feature map of the backbone network. Then, a module for predicting the heat map alone is added after the new module, which uses the fused feature map for predicting the heat map, serving to enhance its own prediction effect.

2.1. Overview. As shown in Figure 1, we design a structure that can utilize shadow features. This structure is used to capture shadow features from the feature map and fuse them into the feature map containing the object, which serves to add features to the object. In our designed model, we use Hourglass [26] as the feature extraction network for this model, unlike CenterNet, we use nonstacked Hourglass. There are two reasons for this choice. First, the stacked Hourglass network is too deep, which will lead to extreme abstraction of the feature layer that will eventually be used as a prediction, and its large cut-off from the feature map in the front part of the network, which will cause difficulties in the design and interpretation of the model. Second, each stacked Hourglass has a different region of interest on the whole picture. Multilevel stacking of Hourglass networks will make each Stack Hourglass network has its own distinct region of interest, i.e., each different Hourglass network needs to add intermediate supervision. And this increases the uncertainty of module design. For these two reasons, we decided to use a nonstacked Hourglass network. The feature map output from this network will be used as the input of the designed module. After the feature map enters the shadow capture module (SCM), it passes through two branches, the first one is the shadow capture branch, we first use three parallel convolutional layers to obtain the initial position parameters of the object in the feature map, and then we capture the features based on the position parameters and our manually designed capture method, using the RoIs (Region of Interests) Align pool to extract the region of interest of the model for shadows, which is stripped from the spatial dimension to the channel dimension. In the second branch, the feature fusion branch, we concatenate the shadow region of interest with the feature map output by hourglass and pass through a convolution layer to complete the fusion of features. Once the fusion is complete, we feed this enhanced feature map into the heat map prediction module added at the end of the model to obtain more accurate object locations and classification scores.

2.2. Shadow Semantic Feature Capture Module. In this section, we will introduce the shadow semantic feature capture module in detail. First, in Figure 2, three parallel predictive convolution layers are set up to obtain the location parameters $\theta(x_c, y_c, w_{\text{obj}}, h_{\text{obj}})$ of the object, where x_c, y_c represents the central point coordinates of the object in the feature map and $w_{\text{obj}}, h_{\text{obj}}$ represents the width and height of the object in the feature map. In view of the imaging characteristics of the experimental data set (see the section Data sets and labels for detail), it can be determined that the shadow of the object in the sonar image generally exists directly above the object. Therefore, we design a capture method for shadow features so that no additional supervision information is required. The rule will eventually capture the shadow features in a selected region on the feature map, and the rule is as follows:

(1) The coordinates of the upper-left corner of the object (x_l, y_l) are obtained from the position parameters of the object, then the height H of the region is $y_l - y_L$, y_L is the vertical coordinate 0 of the upper-left pixel point of the feature map, and the width W is the width w_{obj} of the object.

(2) By observing the shadow images in the data sets, we found that the width of some shadows is slightly larger than the object, and there is also a small skewing of the shadows to the sides in the data sets, so we introduced a width parameter α to adjust the width of this region. The final width of this region should be $\alpha * w_{\text{obj}}$, as shown in Figure 3.

Finally, after the predetermined area is obtained, RoI-Align Pool is used to cut it. For detailed operation and gradient back propagation form, please refer to [27, 28]. The pool can be used to obtain the high response value in the predetermined area and output a feature map with the same size as the original feature map and containing the shaded high response value. Then, the newly generated feature map will be input into the subsequent semantic feature fusion module for feature fusion.

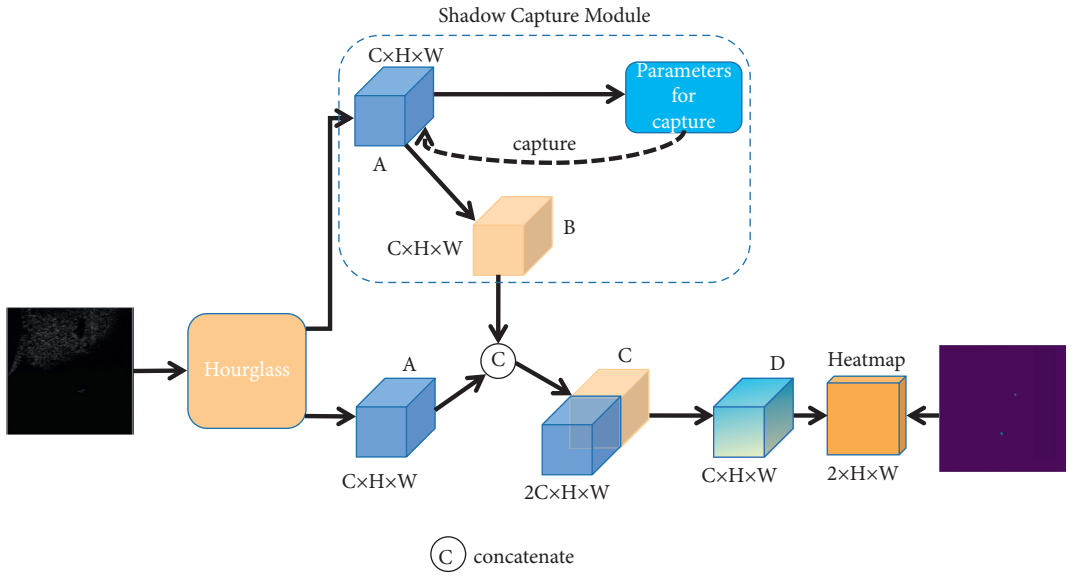


Figure 1: Feature map A generated by Hourglass network is used to predict capture parameters. The obtained parameters are used to capture shadow features in feature map A and then shadow feature map B is obtained. Then, feature figure A and shadow feature figure B are spliced to obtain spliced feature figure C Then, the spliced feature figure C is sent to the fusion module for feature fusion to obtain fusion feature figure D. Finally, the fusion feature figure D is used to predict the final Heatmap.

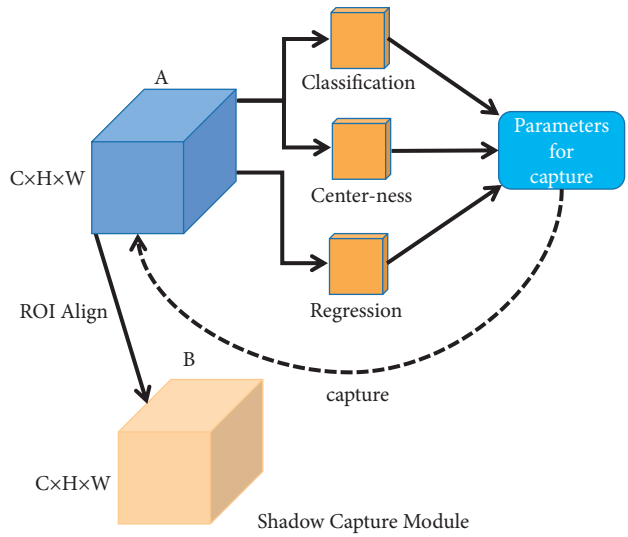


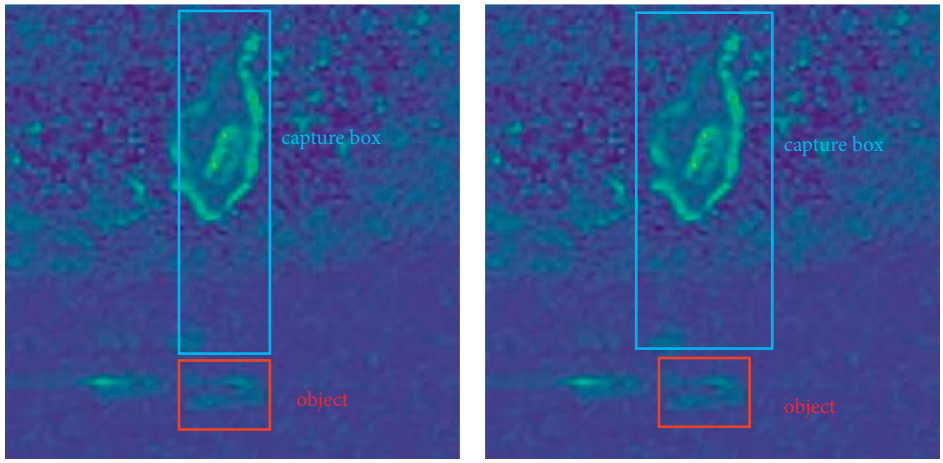
Figure 2: The structure of shadow capture module.

2.3. Semantic Feature Fusion Module. This module takes the shared convolutional feature map output from Hourglass network and the feature map output from the shadow semantic feature capture module containing the high response value of the shadow as input and fuses the two. Finally, the enhanced feature map integrated with the shadow semantic feature is used for the prediction of subsequent heat maps.

As shown in Figure 1, after the shared convolutional feature map is input into the module, it waits for the shadow semantic feature capture module to output the feature map containing the high response value of the shadow and then uses concatenate operation for both and obtains the final feature map C through a fusion function $H(\cdot)$:

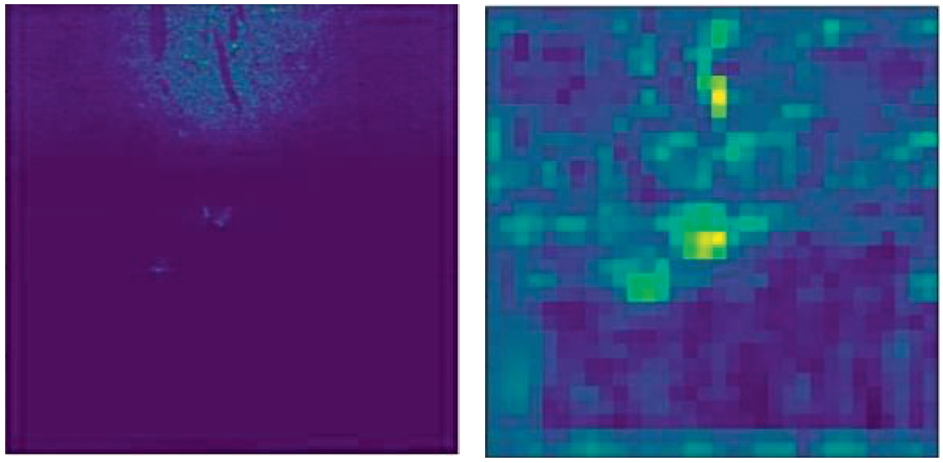
$$C = H(A, B), \tag{1}$$

where $H(\cdot)$ is a mixture of three consecutive operations: 3×3 convolutional layer, Batch Norm [29], and nonlinear activation function ReLU. 3×3 convolutional layer is used to fuse feature map B, containing the shaded high response values, with the feature map of the Hourglass network output A. The Batch Norm of [29] normalizes the values of the output of the convolutional fusion, which can alleviate the Internal Covariate Shift phenomenon (i.e. after each parameter iteration update, the output data of the previous layer will change in data distribution after being computed by this layer of the network, making it



(a) (b)

Figure 3: Capture schematic diagram of shadow features. (a) The capture schematic diagram without adjusting the cutting width and (b) capture schematic diagram after adjusting the cutting width.



(a) (b)

Figure 4: The above two figures are (a) the original figure (processed for convenient observation) and (b) the feature figure in the neural network.The highlighted area in the neural network feature map is the area of interest of the network, as shown in Figure (b). Both shadows and objects are highlighted in the feature map.

difficult for the learning of the next layer), that is inherently present in neural networks [29] and the changes to the model data distribution brought by our use of the concatenate fusion form. The nonlinear activation function ReLU can suppress and activate the fused features with smoothness.

2.4. Interpretability of Modules. In this section, we take a mathematical intuition at the model to better understand why shadow capture module (SCM) is able to incorporate the semantic features of shadows. At the beginning of Section 2, we explained that the effect of model positioning and recognition depends on the response value of pixels in the final feature image. That is, we hope that the final output of the model can have a higher response value to the actual location of the sonar object, and the pixel value of other locations is close to 0 as far as possible. In Figure 4, we found that the feature map also has highlight in shadow areas after visualizing and analyzing. But there are only two goals in the sonar image, if the shadow area still has a high response value, it will increase the fitting pressure of final prediction model and make the model more likely to cause the error detection. Therefore, we hope to use the information of the high-response area in the shadow to improve the response effect of the object area, so as to help the model classification. In order to achieve this goal, it is necessary to strip the high-concern area of the shadow feature from the spatial dimension and then splice the high-concern area to the channel dimension of the feature. After the high-concern areas of shadow features are stripped from the spatial dimension, the values of each pixel in the final feature map should be constituted as follows (taking 1×1 convolution as an example):

$$F_{ji} = W_j * X_{ji}, \tag{2}$$

where j represents different channels and i represents pixel values at different positions on the same channel. W_j is the weight parameter of the filter in the j channel of the input feature map, X_{ji} represents the pixel value of i position in the j channel of the input feature map, and F_{ji} represents the pixel value of i position in the j channel of the output feature map. When the high-response area of the shadow is stripped to the channel dimension, the model can utilize this high-concern area by itself according to the final detection requirements. Since the filters of convolutional neural network in each channel are different, it also means that the parameters of filters are different and are selected by the model, so the filter parameters can be regarded as the model's emphasis on this part of the highlighted region. With the learning of the reverse gradient propagation method, the convolution can learn the situation that is most suitable for the fusion of the feature pixels of each channel.

3. Experiment

In order to evaluate the proposed method, we conducted a comprehensive experiment on the underwater forward-looking sonar data set provided by Pengcheng Laboratory. The experimental results show that our model achieves good performance when the IOU threshold is 0.5–0.8, using PASCAL VOC 2012 as the evaluation standard. In the following sections, we will first describe the data set and the details of the experimental implementation, and then we will present the experimental results of a series of ablation experiments performed on the forward-looking underwater sonar data sets.

3.1. Data Sets and Labels. This data set is the largest and most extensive acoustic image data set in the current industry launched by Pengcheng Laboratory. The data sets have a total of 5000 images, including 3200 training sets, 800 validation sets, and 1000 test sets. The object types include cube, ball, cylinder, human body, tire, circle cage, square cage, metal bucket, and so on. Each image only marks the relevant object, and the shadow does not have any marking information. The data acquisition equipment used for the sonar images was the Tritech Gemini 1200I multibeam forward-looking sonar. The detection beam emitted by the equipment is horizontally divided into several fan-shaped beams with vertical opening angle φ, each beam irradiates the object as shown in Figure 5 and forms a set of distance intensity information, and the echo intensity information of all the beams is arranged according to position relation to form sonar image, as shown in Figure 6. The number of beams corresponds to the number of horizontal pixels of the image. When the horizontal angle θ is constant, the more the beams, the higher the angular resolution, the amount of echo data collected by each beam corresponds to the number of vertical pixels of the image, the larger the amount of data, the higher the distance resolution. The sonar data set is Cartesian (after a rectangular coordinate system) so that the image appears as a rectangle rather than a fan. As shown in Figure 6, the pixel (x, y) in the sonar image represents the acoustic reflection intensity information at direction $\theta = (W/2 - x) \cdot \varphi/W$ and range $r = (H - y) \cdot R/H$ in the polar coordinate system. φ and R, respectively, represent the horizontal opening angle and slant range of the forward-looking sonar, W and H, respectively, represent the horizontal and vertical dimensions of the image.

3.2. Experimental Details. We implemented our method based on Pytorch. After many experiments and comprehensive consideration, we adopted the following settings to train the sonar data set from scratch. The size of input resolution is fixed at 512×512, the optimizer uses Adam, the basic learning rate is set at 0.001, the number of training rounds is 300, the training strategy of learning rate fixed step size reduction is adopted, each time the reduction is 1/10, and the number of decreased rounds is 120,200,260, respectively. In order to enhance the diversity of data, we used conventional data processing methods (vertical flip, horizontal flip, etc.) for the images.

3.3. Loss Function. The training loss function of our model consists of two parts. The first part is the training loss in the stage of shadow location prediction. In this stage, not only the heatmap (heatmap1) was predicted but also the length and width of the object and the offset of the object center point. Therefore, the loss values in this stage are as follows: loss of preliminary heatmap (heatmap1), object length and width prediction loss, and object center offset loss. In the shadow prediction stage, the loss value of the predicted heatmap1 L_{k1} was calculated by focal loss [30], both the loss value of the predicted length and width L_{size} and the loss value of the object center offset L_{off} were calculated by L1 loss, and λ_{size} and λ_{off} were both set as 0.1.

$$L_{\text{det1}} = L_{k1} + \lambda_{\text{size}} L_{\text{size}} + \lambda_{\text{off}} L_{\text{off}}. \tag{3}$$

The other part is the loss function that predicts the final heatmap, where L_{k2} is the loss of heatmap2.

$$L_{\text{det2}} = L_{k2}. \tag{4}$$

The loss function of the whole model can be obtained by adding the losses of the above two parts, and the specific calculation formula is as follows:

$$L_{sum} = L_{\text{det1}} + L_{\text{det2}}. \tag{5}$$

3.4. Comparison Experiments

3.4.1. The Parameter α of Cutting Width. Since the transverse position of the shadow in this data sets is offset to a certain extent relative to the object, we design a parameter α for the multiples of the cutting width and take different values of α for experimental comparison. The higher the average accuracy (mAP) is, the more consistent the cut width is with the shadow width of the data sets, and the more

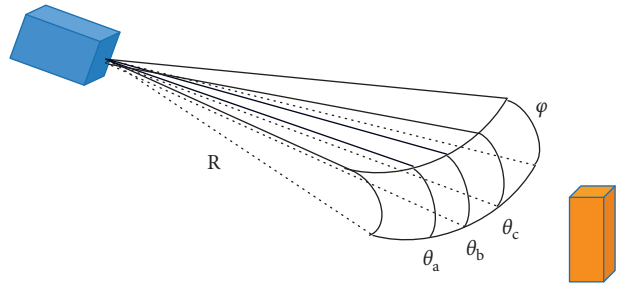


Figure 5: The schematic diagram of the detection beam.

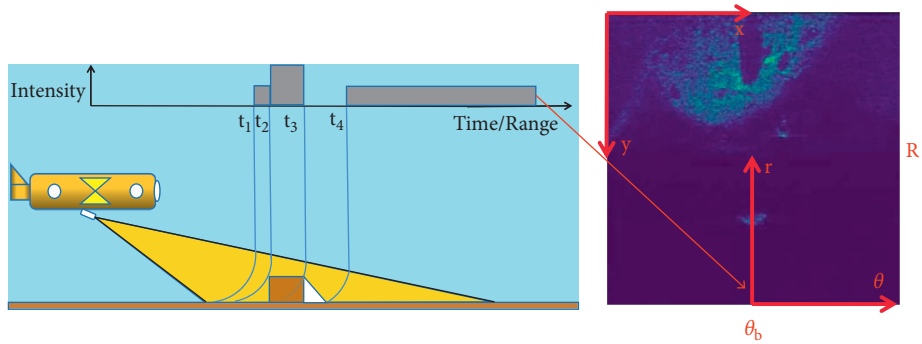


Figure 6: The schematic diagram of underwater forward-looking sonar image.

Table 1: In PASCAL VOC2012 standard, the average accuracy of each width multiple parameter α at different IOU.

α	mAP@0.5 (%)	mAP@0.6 (%)	mAP@0.7 (%)	mAP@0.8 (%)
1	98.12	93.19	71.28	21.62
1.25	97.09	**94.23**	73.95	22.33
1.5	97.51	93.39	**75.73**	**25.47**
1.75	97.11	91.76	70.13	19.83
2	**98.12**	92.24	73.75	21.01

Table 2: Under PASCAL VOC2012 standard, the average accuracy of each model under different IOU.

Method	mAP@0.5 (%)	mAP@0.6 (%)	mAP@0.7 (%)	mAP@0.8 (%)
YOLOv3 [31]	93.80	81.50	56.40	14.40
SSD [32]	91.31	82.75	58.61	19.62
CenterNet (hourglass) [25]	97.42	92.16	69.61	24.97
CenterNet (ResNet-50)	96.74	90.07	65.16	16.24
RefineDet [33]	94.64	88.38	66.62	18.76
FCOS [34]	90.72	87.95	70.70	28.63
Ours	**97.51**	**93.39**	**75.73**	**25.47**

complete the cut shadow features are. The experimental results are shown in Table 1.

3.4.2. Model Comparison. Comparison is made between our mothed and all classical object detection motheds on the same underwater forward-looking sonar data sets. Table 2 shows the comparison of different IOU under PASCAL VOC2012 evaluation standard. We can find that our model has obvious advantages in detection accuracy, by comparing different models. According to the comparison of visual images in Figure 7, it can also be seen that the detection performance of this model is improved compared with CenterNet.

3.4.3. Accuracy of each category. Under Pascal voc2012 evaluation standard, the accuracy performance of our model and classical object detection models on eight classes of objects in sonar data set is compared when IOU is 0.7. Because our structure integrates shadow information into the feature map and enhances the performance of the features of the object on the feature map, it can be seen from

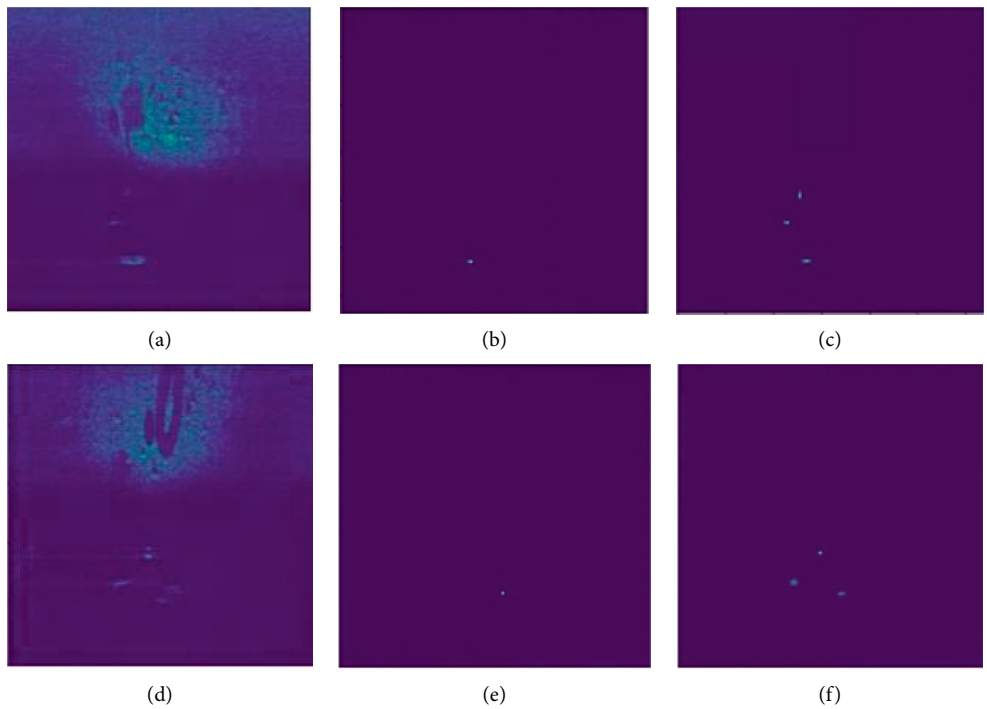


Figure 7: We visualize the heat map predicted by CenterNet and the heat map predicted after adding SCM. A row is a group of comparison pictures, (a, d) are the original pictures (processed for convenience of observation), (b, e) are the heat maps predicted by CenterNet, and (c, f) are the heat maps predicted after adding SCM. It is obvious that the object predicted by the model with SCM is more comprehensive and more accurate.

Table 3: In PASCAL VOC2012 standard, IOU is 0.7, and the accuracy of each model is different.

Methods	mAP@0.7 (%)	Ball (%)	Cylinder (%)	Square cage (%)	Cube (%)	Circle cage (%)	Human body (%)	Metal bucket (%)	Tire (%)
SSD	58.61	71.74	25.13	43.09	74.07	68.54	64.39	63.05	58.91
YOLOv3	56.40	55.20	48.80	49.00	66.40	58.30	59.80	60.80	52.50
RefineDet	66.62	72.56	61.94	54.66	77.88	78.28	57.61	70.89	59.12
CenterNet (hourglass)	69.61	78.50	67.94	64.86	77.89	74.70	67.88	64.60	60.47
CenterNet (ResNet-50)	65.16	73.79	51.30	67.10	75.55	69.28	56.30	71.52	56.24
FCOS	70.70	73.84	62.30	68.59	74.95	78.43	63.17	**76.64**	67.66
Our	**75.73**	**83.12**	**69.47**	**70.28**	**84.20**	**80.82**	**72.73**	71.47	**73.78**

Table 4: Comparison the number of parameters between CenterNet and our model.

Model	Num of pram
CenterNet	95.436 M
Ours	98.390 M
Growth rates	3.09%

Table 3 that the detection performance of this model is almost the best in all categories.

3.4.4. Analyzing the Number of Parameters of SCM. As shown in Table 4, after comparing the number of parameters between CenterNet and our model, we found that SCM only adds 2.954 M of parameters, which is only 3.09% of CenterNet, so this module is very lightweight.

4. Conclusions

In this paper, we propose a structure for capturing shadow features and fusing them into feature maps. It makes use of the correlation between shadow features and object features, as well as the difference of shadow features between different categories of objects, so as to increase the difference of features between categories. For objects with more obvious features, the recognition effect of the network is often better. Combined with our experimental results, it can be shown that the fusion of shadow and object features can indeed enhance the role of network recognition to a certain extent. In addition, there is still something worth exploring about the fusion mode of shadow and object features. How to better combine shadow features and object features will be studied in the future.

Data Availability

The data sets used and analysed during the current study are available from the corresponding author upon reasonable request.

Disclosure

Taowen Xiao and Zijian Cai are the co-first authors of the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest associated with the manuscript.

Acknowledgments

Taowen Xiao, Zijian Cai, Cong Lin, and Qiong Chen contributed equally to this work. This work was supported by the National Natural Science Foundation of China under Grant 62072121 and Natural Science Foundation of Guangdong Province 2021A1515011847.

References

[1] Q. Chen, M. Huang, and H. Wang, "A feature discretization method for classification of high-resolution remote sensing images in coastal areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8584–8598, 2021.

[2] Q. Chen, M. Huang, H. Wang, and G. Xu, "A feature discretization method based on fuzzy rough sets for high-resolution remote sensing big data under linear spectral model," *IEEE Transactions on Fuzzy Systems*, vol. 59, no. 10, pp. 8584–8598, 2021.

[3] H. Wang, X. Li, R. H. Jhaveri et al., "Sparse bayesian learning based channel estimation in FBMC/OQAM industrial IoT networks," *Computer Communications*, vol. 176, pp. 40–45, 2021.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*, pp. 886–893, San Diego, CA, USA, June 2005.

[5] M. Zhu, Y. Song, J. Guo et al., "PCA and kernel-based extreme learning machine for side-scan sonar image classifification," in *Proceedings of the IEEE Underwater Technology (UT)*, pp. 21–24, Busan, South Korea, February 2017.

[6] V. Myers and J. Fawcett, "A template matching procedure for automatic target recognition in synthetic aperture sonar imagery," *IEEE Signal Processing Letters*, vol. 17, no. 7, pp. 683–686, 2010.

[7] R. Fandos, A. M. Zoubir, and K. Siantidis, "Unified design of a feature based ADAC system for mine hunting using synthetic aperture sonar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2413–2426, 2014.

[8] Y. LeCun, B. Boser, J. S. Denker et al., "Backprop agation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classifification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[11] S. Wang, B. Kang, J. Ma et al., "A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)," *European Radiology*, vol. 31, 2021.

[12] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, vol. 121, Article ID 103792, 2020.

[13] M. Valdenegro-Toro, "Object recognition in forward-looking sonar images with convolutional neural networks," in *Proceedings of the OCEANS 2016 MTS/IEEE Monterey*, September 2016.

[14] M. Valdenegro-Toro, "End-to-end object detection and recognition in forward-looking sonar images with convolutional neural networks," in *Proceedings of the 2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*, November 2016.

[15] D. P. Williams, "Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks," in *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, December 2016.

[16] A. Galusha, J. Dale, J. M. Keller, and A. Zare, "Deep convolutional neural network target classification for underwater synthetic aperture sonar imagery," in *Proceedings of the Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV*, vol. 11012, April 2019.

[17] K. Breistein, "*Applying machine learning using custom trained convolutional neural networks on subsea object detection and classification*," M.S. thesis, NTNU, Trondheim, Norway, 2019.

[18] G. Neves, M. Ruiz, J. Fontinele, and L. Oliveira, "Rotated object detection with forward-looking sonar in underwater applications," *Expert Systems with Applications*, vol. 140, Article ID 112870, 2020.

[19] L. Zacchini, M. Franchi, V. Manzari et al., "Forward-looking sonar CNN-based automatic target recognition: an experimental campaign with FeelHippo AUV," in *Proceedings of the 2020 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV) (50043)*IEEE, St. Johns, Canada, 2020.

[20] Z. Fan, W. Xia, X. Liu, and H. Li, "Detection and segmentation of underwater objects from forward-looking sonar based on a modified Mask RCNN," *Signal, Image and Video Processing*, vol. 15, pp. 1–9, 2021.

[21] S. Reed, Y. Petillot, and J. Bell, "An automatic approach to the detection and extraction of mine features in sidescan sonar," *IEEE Journal of Oceanic Engineering*, vol. 28, pp. 90–105, 2003.

[22] E. Sang, Z. Shen, C. Fan, and Y. Li, "Sonar image segmentation based on implicit active contours," in *Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 4, November 2009.

[23] V. Myers and J. Fawcett, "A template matching procedure for automatic target recognition in synthetic aperture sonar imagery," *IEEE Signal Processing Letters*, vol. 17, no. 7, pp. 683–686, 2010.

[24] V. Myers and D. P. Williams, "Adaptive multiview target classification in synthetic aperture sonar images using a partially observable Markov decision process," *IEEE Journal of Oceanic Engineering*, vol. 37, pp. 45–55, 2011.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, https://arxiv.org/abs/1904.07850.

[26] A. Newell, K. Yang, and D. Jia, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, October 2016.

[27] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: hard positive generation via adversary for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[29] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning. PMLR*, Lille, France, July 2015.

[30] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[31] J. Redmon and F. Ali, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[32] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, October 2016.

[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, July 2018.

[34] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, October 2019.