EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Side-scan sonar underwater target segmentation using the BHP-UNet

Yulin Tang[1*], Liming Wang[1], Houpu Li[1] and Shaofeng Bian[1]

*Correspondence:
yltanghg@163.com

[1] College of Electrical Engineering, Naval University of Engineering, Wuhan 430033, Hubei, China

Abstract

Although target detection algorithms based on deep learning have achieved good results in the detection of side-scan sonar underwater targets, their false and missed detection rates are high for multiple densely arranged and overlapping underwater targets. To address this problem, a side-scan sonar underwater target segmentation model based on the blended hybrid dilated convolution and pyramid split attention UNet (BHP-UNet) algorithm is proposed in this paper. First, the blended hybrid dilated convolution module is adopted to improve the ability of the model to learn deep semantics and shallow features while improving the receptive field. Second, the pyramid split attention module is introduced to establish a long-term dependency between global and local information while processing multi-scale spatial features. Three sets of experimental results show that the BHP-UNet model proposed in this paper has better segmentation performance than the conventional fully convolutional network, UNet, and DeepLabv3+ models, and it is able to segment dense and overlapping targets to a certain extent. The proposed model will have significance as a guide for practical applications.

Keywords: BHP-UNet algorithm, BHD module, PSA module, Target segmentation, Side-scan sonar images, Deep learning

1 Introduction

Underwater target detection is an important component of navigation safety, marine resources investigation, obstacle verification, and underwater search and rescue. The most commonly used seabed geomorphology and underwater target detection device is side-scan sonar, which has played an important role in the detection of seabed targets. Side-scan sonar is predominantly based on the echo detection principle for the purpose of underwater target detection [1]. It employs a carefully designed transducer array that emits pulsed ultrasonic waves at a specific tilt angle, facilitating directional transmission characterized by a broad vertical beamwidth and a narrow horizontal beamwidth. These emitted waves propagate toward the seafloor or underwater targets, interacting with them through reflection and scattering phenomena. The receiving transducer array effectively captures the resulting reflections and scattered waves from the seafloor. Afterward, the received signals are amplified, processed, and recorded to generate a visual representation of the seafloor on a monitor.

Conventional side-scan sonar images are manually interpreted, which is inefficient, time consuming, and strongly dependent on human experience [2]. Machine-learning methods, which are based on extracting the target texture, grayscale information, edge information, and other features, have problems such as overfitting, poor generalization ability, and poor robustness in terms of feature extraction [3–9]. In contrast, the side-scan sonar seabed underwater target detection techniques based on deep learning are becoming the research focus of international and Chinese researchers [10–17]. For instance [18], proposed a convolutional neural network transfer-learning recognition method using an improved VGG-16 as the framework. The method was able to perform automatic image recognition of side-scan sonar seabed shipwrecks and achieve significantly better accuracy and efficiency than the classical machine-learning SVM algorithm. The authors of [19] automatically detected side-scan sonar seabed shipwreck targets using the faster RCNN model. However, the model has problems such as a complex structure, and low efficiency in training and detection. The authors of [20] proposed a side-scan sonar shipwreck target detection method based on transfer learning with an improved YOLOv3 model, which increased the training and detection efficiency to a certain extent, but there are still problems, such as a high missed detection rate for small targets and a detection speed that does not meet real-time requirements. To meet practical engineering application requirements [21], proposed an improved YOLOv5a model based on YOLOv3, which achieved good results in detection accuracy and efficiency for small-scale targets. However, its detection performance in complex sea conditions needs to be improved. To this end [22], proposed a lightweight DETR-YOLO model for side-scan sonar seabed shipwreck detection, which improved the global scene understanding and detection in complex marine environments while meeting lightweight engineering deployment requirements. The above algorithms have achieved good results in the detection of side-scan sonar seabed shipwreck targets, but they have high false and missed detection rates for densely arranged, overlapping, and complex multiple targets, and cannot meet the requirements of actual engineering tasks.

In addition to target detection, the core research in the field of computer vision also includes semantic segmentation, which can be understood as the assignment of semantic labels to each pixel of the image and subsequent division of the image into several areas with different semantic identities based on the semantic unit labels. It is a pixel-level, dense prediction task. In recent years, the emergence of deep convolutional neural networks has greatly facilitated the development of semantic segmentation, and has become increasingly applied in the fields of intelligent security, autonomous driving, satellite remote sensing, medical image processing, biometric recognition, virtual reality, and augmented reality. Long et al. [23] proposed an image semantic segmentation method based on fully convolutional networks (FCNs) in 2014, which became a classical method for semantic segmentation. In this method, the main fully connected layer is replaced with a convolutional layer, and the semantic image segmentation problem is transformed into a classification problem for each pixel. In the same year, a fully convolutional network structure was proposed called UNet [24], which is primarily used for medical image segmentation. The UNet model adopts the encoder–decoder structure, which has been widely used in the field of biomedical image segmentation. The DeepLab semantic segmentation series of models consists of DeepLabv1-DeepLabv3+ [25–28].

Specifically, the DeepLabv3+ model is a semantic segmentation algorithm with excellent overall performance because it combines the advantages of the encoder–decoder and atrous spatial pyramid pooling, which expands the receptive field without changing the resolution and combines features at different scales. However, because of its complex model structure and many parameters, DeepLabv3+ requires a large number of data samples to train the model to achieve high segmentation performance. Hence, it is most suitable for semantic segmentation tasks with large data sample sizes and high target complexity.

By contrast, side-scan sonar seabed underwater target images (the seabed shipwreck is used as a typical object in this article) have the following two main features [29–31]: (1) the semantics of the seabed underwater target image are relatively simple and the structure is relatively fixed. The seabed underwater target is relatively fixed in sonar mapping, and hence deep semantic information and low-level features are very important. (2) The amount of available data is limited. In reality, there are relatively few seabed underwater target, and it is difficult and expensive to obtain side-scan sonar data from it. Therefore, the semantic segmentation model should not have an excessively complex structure, and the number of parameters should not be too large, otherwise the model will be prone to overfitting and the segmentation results will be poor.

Given the features of side-scan sonar seabed underwater target images and with the aim of solving the problems of high false and missed detection rates for multiple seabed underwater targets that are densely arranged and overlapping, a side-scan sonar seabed underwater target segmentation model based on the blended hybrid dilated convolution and pyramid split attention UNet (BHP-UNet) model is proposed in this paper. The proposed method is inspired by the UNet segmentation method. The primary objective is to improve the performance of underwater target segmentation under complex marine conditions and offer practical guidance for real-world applications. The key contributions of this research can be summarized as follows:

- *BHD module* A specially designed module, known as the BHD module, enhances the model's receptive field while effectively integrating deep semantic and shallow features. This module improves the model's learning capacity by combining multi-scale information, including deep semantics and shallow features.
- *PSA module* The PSA module is introduced to handle spatial features at multiple scales and establish long-term dependencies between global and local information. This enables the model to better comprehend the entire scene and enhance its contextual understanding.
- *Integration of BHD module, PSA module, and UNet model* By leveraging the characteristics of side-scan sonar images, the study innovatively combines the BHD module, PSA module, and UNet model to create the BHP-UNet model. This integrated model aims to improve the segmentation performance of shipwreck targets by addressing their dense arrangement and overlap.

To evaluate the effectiveness of the proposed model, a comprehensive analysis is conducted, including real-world offshore trials in Zhoushan, model comparison experiments, ablation experiments, and underwater simulation experiments. The evaluation

aims to overcome the challenges in shipwreck target detection using side-scan sonar, specifically the issues of high missed detection rates and false alarm rates.

The paper is organized into four sections. The introduction provides the necessary background and outlines the objectives of the study. Section 2 introduces the BHP-UNet model, detailing its composition and underlying principles. In Sect. 3, the proposed method is thoroughly evaluated and extensively discussed. Finally, Sect. 4 concludes the paper by summarizing the key findings, highlighting the contributions, and outlining future research directions.

2 Proposed BHP-UNet model

The BHP-UNet model structure, in which an encoder–decoder architecture was adopted, is shown in Fig. 1. The left half of the encoder extracts multi-scale features through convolution and pooling operations. It includes four BHD modules to expand the receptive field while maintaining the resolution and spatial hierarchy information of the image, where receptive field is the spatial extent influencing neuron activation. It varies with kernel size and network depth. Larger fields capture global context, while smaller ones focus on local details. They play a key role in information integration across scales. After the BHD operations, three average pooling operations are performed to extract multi-scale features while reducing computation and the memory used in the graphics processing unit.

In the decoder, shown on the right of Fig. 1, the resolution is restored through three upsampling (Upconv) layers, and the final feature map is generated using feature concatenation (Concat) and convolution operations. The PSA module, which integrates multi-scale spatial information and cross-channel attention into each segmented feature, is introduced before feature concatenation. This is done to improve the interaction between local channel attention and global channel attention. Shallow and deep semantic features are combined using feature concatenation.

2.1 BHD module

Considering the overall correlation of the side-scan sonar underwater target image, the importance of abstract image features, and the differences between each image sample, expanding the receptive field of the network and combining heterogeneous receptive fields improves the performance of feature extraction in the model. Although a pooling operation can effectively expand the receptive field and reduce the complexity of the model, it will reduce the resolution of the image and lead to the loss of image structure and spatial hierarchical information, which causes pixel-level segmentation to fail. Dilated convolution [32] can expand the receptive field while maintaining the resolution of the image without the need to introduce additional parameters, but continuous dilated convolution will cause grid artifacts, which degrades the segmentation results of small-scale targets. To address the problems of the pooling operation and dilated convolution, the BHD module, which performs multi-scale dilated convolution module with expansion rates of 1, 2, and 5 and then combines heterogeneous receptive fields is adopted. The convolution diagram is shown in Fig. 2.

The equation for calculating the size of the expanded dilated convolution kernel is as follows:

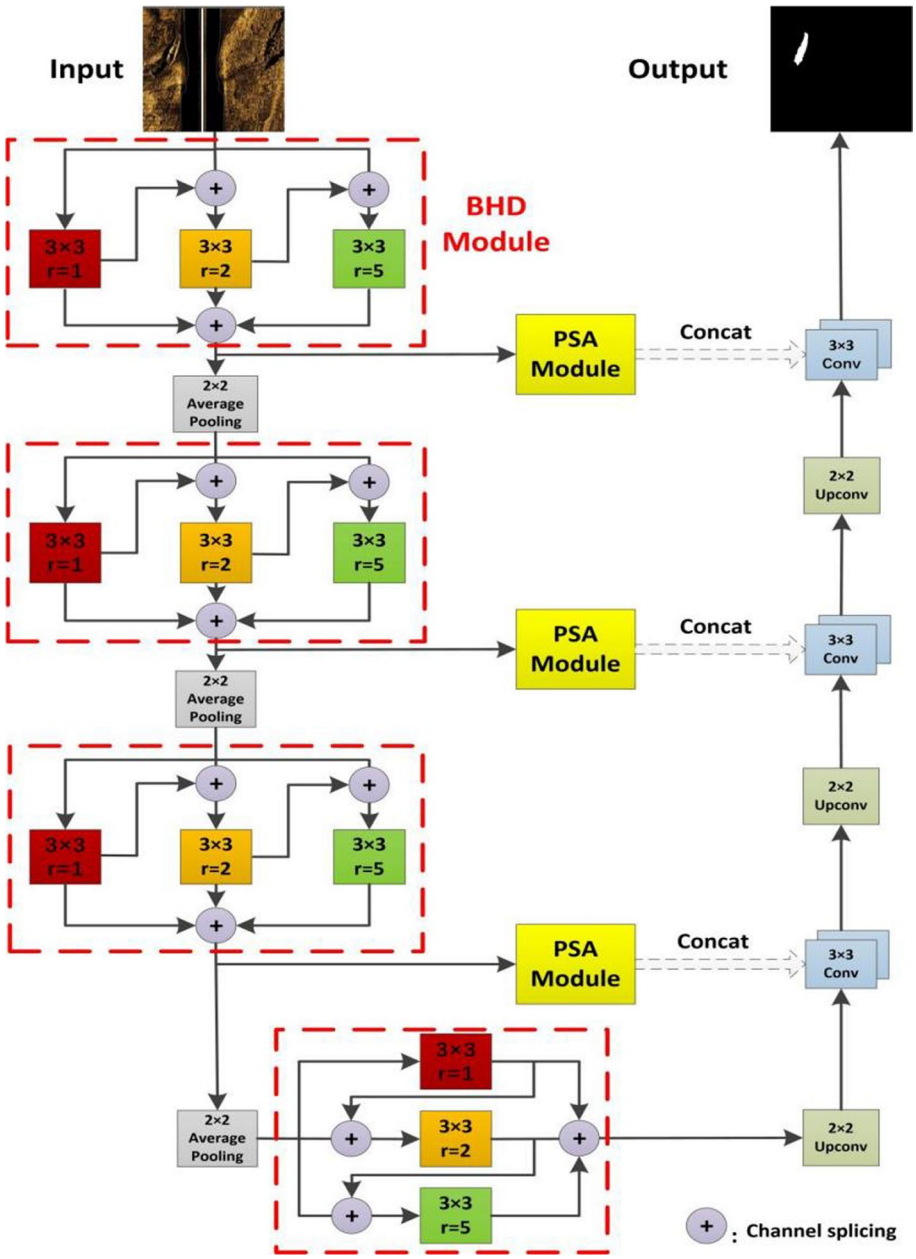

Fig. 1 BHP-UNet model structure

$$K_{\mathrm{r}} = K + (K-1)(r-1), \tag{1}$$

where K is the size of the original convolution kernel and r is the convolution expansion rate.

The advantage of dilated convolution is that it expands the receptive field without introducing additional parameters, but using dilated convolutions with the same expansion rate leads to grid artifacts during the convolution operation. However, if only one dilated convolution with a large expansion rate is used, it may only segment some large targets, and the segmentation of small targets will be poor. To this end,

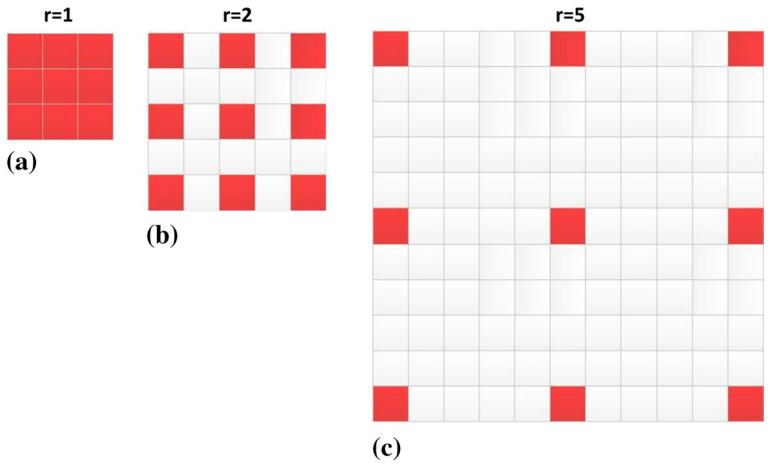

Fig. 2 Dilated convolution. **a** Shows a standard 3 × 3 convolution kernel, and **b**, **c** show convolution kernels that have been expanded at different rates. **b** Shows a dilated convolution with an expansion rate of 2 ($r = 2$), namely $r - 1$ zeros are added between the elements of a standard 3 × 3 convolution kernel, and its receptive field is equivalent to that of a 7 × 7 convolution kernel. **c** Shows a dilated convolution with an expansion rate of 5 ($r = 5$). In this case, four zeros are added between the elements of a standard 3 × 3 convolution kernel, and its receptive field is equivalent to that of an 11 × 11 convolution kernel

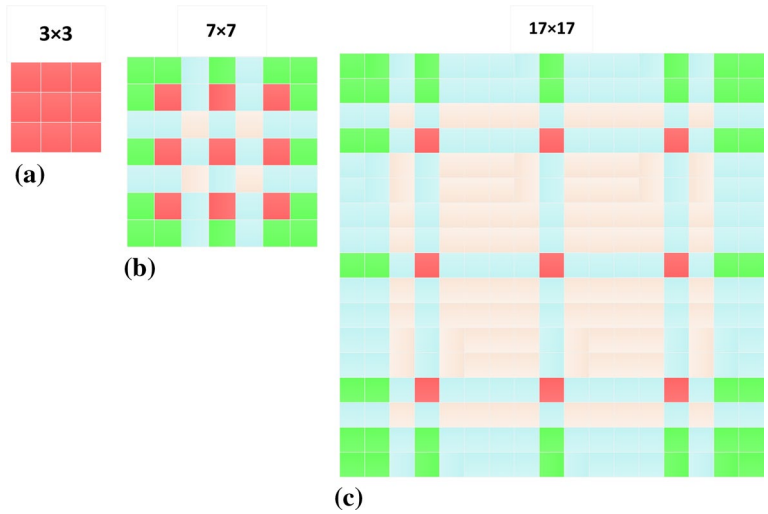

Fig. 3 Hybrid dilated convolution. **a** represents the receptive field of $r = 1$; **b** represents the receptive field of r = 2; **c** represents the receptive field of $r = 5$

hybrid dilated convolution is proposed in this paper to avoid introducing grid artifacts. It is illustrated in Fig. 3.

Figure 3 shows the changes in the size of the receptive field in each layer and the number of times each pixel in the receptive field is learned in hybrid dilated convolution, where yellow refers to four learning iterations, blue refers to two learning iterations, green refers to one learning iteration, and red denotes the magnitude of the convolution expansion rate of each layer. The receptive field of each layer is calculated as follows:

$$Rf_n = Rf_{n-1} + (K_r - 1)\prod_{i=1}^{n-1} s_i, \tag{2}$$

where Rf_n is the receptive field of the local layer, Rf_{n-1} is the receptive field of the previous layer, s_i is the step size of the convolution on the ith layer, and K is the size of the convolution kernel. In this paper, the step size (stride) is 1 and the size of the convolution kernel is 3. According to Eq. (2), the receptive field of the first layer is 3×3, that of the second layer is 7×7, and that of the third layer is 17×17. The hybrid dilated convolution strategy not only greatly expands the receptive field while avoiding grid artifacts, but also focuses on learning the important parts of the receptive field.

Under normal circumstances, segmenting the image using the network can identify one or more regions with a high-level feature with a high response that can be used to correctly segment the target. However, when there is a target that is large in size, although the conventional convolution kernel can obtain accurate positioning information and highlight the characteristic features of the target, much of the regional information related to the target will be missed, which makes it difficult to generate comprehensive, complete, and dense target positioning, thereby limiting the overall segmentation performance of the model. Therefore, hybrid dilated convolution is used to propose a multi-scale hybrid dilated convolution combination strategy in this paper. This strategy expands the receptive field by changing the expansion rate of the convolution kernel at multiple scales so that a target area with a low response can be better recognized by perceiving the semantic features with a high response in the surrounding region. Moreover, it transfers the knowledge of an unusual area containing distinguishing features to the adjacent target area so that the feature information of the high-response part of the target can be propagated to the adjacent target area at multiple scales. Finally, the target recognition and positioning maps generated under different expansion rates are combined to enable dense and accurate target recognition and positioning, which improves the recognition ability of the segmentation model. The process of multi-scale hybrid dilated convolution blending is illustrated in Fig. 4.

As shown in Fig. 4, only a small area with the most distinctive characteristics near the seabed underwater target center was located using the 3×3 convolution kernel with an expansion rate of 1. After the expansion rate was increased from 1 to 2 and the knowledge of the previous layer was combined with that of the next layer, the area near the seabed underwater target could be perceived. Furthermore, the low-response areas such as the seabed underwater gunwale and outline around the seabed underwater target were perceived after the expansion rate was further increased to 5 and the upper layer was further combined. Therefore, multi-scale hybrid dilated convolutional blending can locate and learn more complementary feature areas while focusing on learning the high-response feature areas, thereby comprehensively improving the segmentation ability of the model.

As Fig. 4 shows, however, some false positive (FP) nontarget feature areas may be incorrectly enlarged by dilated convolution at high expansion rates (e.g., $r=5$ partial areas). Hence, an anti-noise blending strategy was proposed to solve this problem that is expressed as follows:

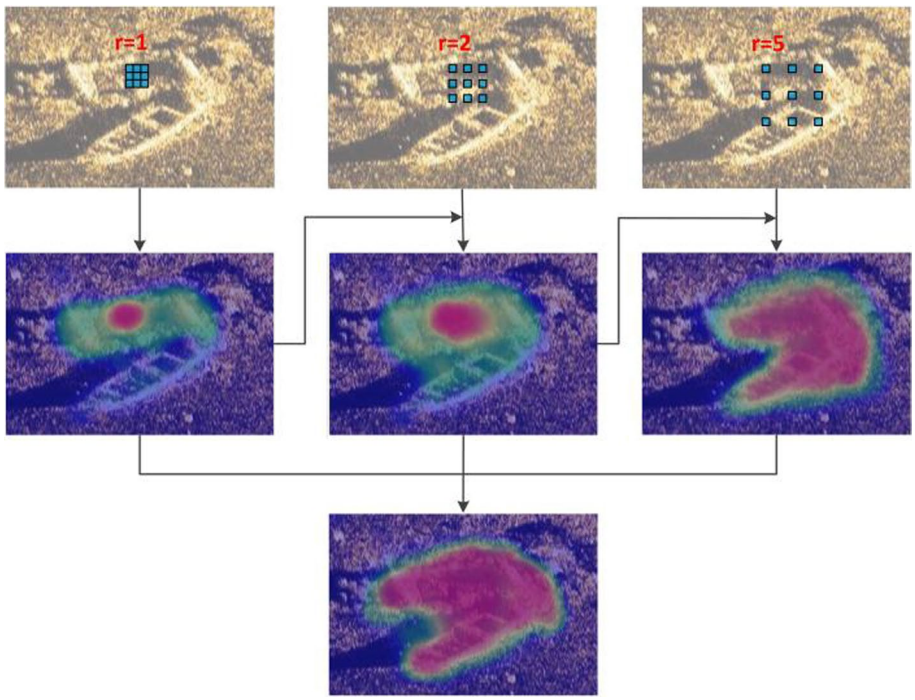

Fig. 4 Heatmap showing the multi-scale hybrid of dilated convolutions for underwater targets

$$L = L_1 + \frac{1}{n_d}\sum_{i=1}^{r} L_i, \tag{3}$$

where L refers to the final receptive field, L_1 and L_i denote the receptive fields when dilated convolutions with expansion rates of 1 and i are used, respectively, and n_d is the number of dilated convolutions. The final blending diagram at the bottom of Fig. 4 shows that the anti-noise blending strategy has effectively suppressed the response in the areas that are not related to the target features included in the expanded receptive field. Moreover, it has combined the localization information in the areas generated by different expansion rates into a complete localization map with the target feature area highlighted in a better way.

2.2 PSA module

The conventional attention mechanism used in SENet [33] only considers channel attention and ignores spatial attention. CBAM [34] considers channel attention and spatial attention, but spatial information at different scales is not captured to enrich the feature space, and spatial attention only considers the information in the local area such that long-distance dependency cannot be established. Although PyConv [35] addresses this problem, it is a complex model and requires a large amount of computation. Therefore, the efficiency of feature information usage at different scales of seabed underwater images is not high, channel attention can only effectively capture local features, and it is difficult to establish global long-term dependencies. Because of the additional computation caused by the BHD module introduced in Sect. 2.1, a lightweight and effective PSA module is adopted, as shown in Fig. 5.

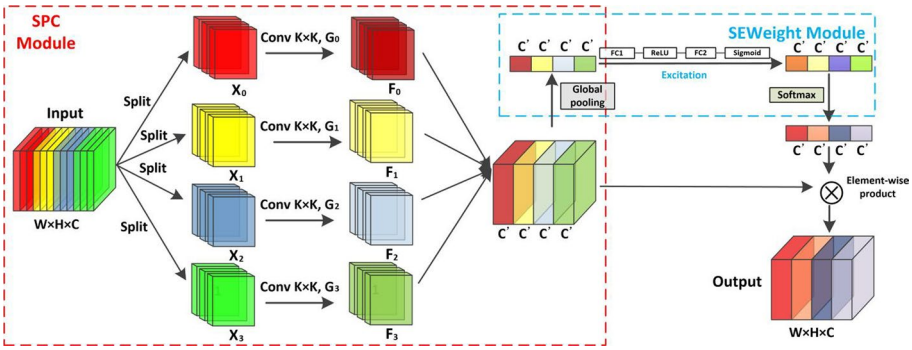

Fig. 5 PSA module structure

The PSA module is divided into three parts: the split-and-concatenate (SPC) module, SEWeight module, and Fscale. The specific operations include the following four steps: first, channels are sliced in the SPC module to obtain a spatial scale feature map. Second, the SEWeight module is used to obtain channel attention vectors at different scales. Third, the softmax function is used to recalibrate and allocate the attention vectors of the channels at different scales and obtain the multi-scale channel attention weights of the new feature map. Finally, the new attention weights and original feature map are multiplied element-wise to obtain the final multi-scale feature attention-weighted response map.

In the PSA module, the multi-scale features are extracted by the SPC module, and the input feature map X is first divided into S parts X_0, X_1,..., X_{S-1}. For each segmented part, there are $C' = \frac{C}{S}$ common channels, and the first feature map i is denoted as $X_i \in R^{C' \times H \times W}, i = 0, 1, 2, \ldots, S-1$. After segmentation has been performed, the spatial information in the feature map of each channel is extracted by processing the input tensors at multiple scales in parallel, and feature maps with different spatial resolutions and depths are generated using multi-scale convolution kernels in a pyramid structure. The purpose of the splitting process in the SPC module is to enable the model to independently learn multi-scale spatial information and establish cross-channel interaction in a local way for each split part. However, as the size of the convolution kernel increases, this increases the number of parameters for the entire model substantially. Therefore, to process input tensors at different scales without increasing the computational cost, a grouped convolution method is introduced and applied to the convolution kernels in parallel. In this method, the relationship between the size of the multi-scale convolution kernel and the size of each group is as follows:

$$G = 2^{\frac{K-1}{2}}, \tag{4}$$

where K is the size of the convolution kernel and G is the size of each group.

Therefore, the multi-scale feature map function is as follows:

$$F_i = \mathrm{Conv}(K_i \times K_i, G_i)(X_i), \quad i = 0, 1, 2, \ldots, S-1 \tag{5}$$

where $K_i = 2 \times (i+1) + 1$ and $G_i = 2^{\frac{K_i-1}{2}}$. The values of i=0, 1, 2, 3, $F_i \in R^{C' \times H \times W}$ used in this paper refer to the feature maps of different scales and were selected

experimentally. Finally, the feature map after multi-scale blending is obtained by concatenation as follows:

$$F = \mathrm{Concat}([F_0, F_1, \cdots, F_{S-1}]) \tag{6}$$

where $F_i \in R^{C' \times H \times W}$ is the final multi-scale feature map and Concat refers to feature concatenation in the channel dimension.

After the multi-scale features have been extracted by the SPC module, the channel attention weight information is extracted using the multi-scale feature map of the SEWeight module. The SEWeight module is divided into two parts, a squeeze part and an excitation part. The squeeze part compresses the corresponding feature map in one dimension through global pooling, that is, the $W \times H \times C'$ feature map is compressed to $1 \times 1 \times C'$ using

$$Sq_i = \frac{1}{W \times H} \sum_{j=1}^{W} \sum_{k=1}^{H} u_i(j,k) \tag{7}$$

where W and H refer to the width and height of the feature map, respectively, C is the number of the channels, $u_i(j,k)$ denotes the (j,k)th element in the ith feature map, $i \in C'$. After the global features have been obtained through the squeeze operation, the relationship between the channels is extracted as follows through the excitation operation:

$$Ex = \sigma\big(g(z, W)\big) = \sigma(W_2 \delta(W_1, z)), \tag{8}$$

In the excitation operation, a gating mechanism based on the sigmoid function is adopted, in which the number of the channels is reduced to $1/r$ of the original number using parameter W_1 using fully connected layer FC_1. After activation by a ReLU function, the number of channels is restored to the original number using parameter W_2 through fully connected layer FC_2. Finally, the weights of each channel are generated using a sigmoid activation function. The dimensionality reduction ratio used in this study is $r = 16$, which was determined experimentally.

Finally, the generated weight values are applied to the corresponding feature channel F_i using the scale operation to obtain the final output Z_i as follows:

$$Z_i = F_{\mathrm{scale}}(u_i) = u_i \times Sq_i, \tag{9}$$

The SEWeight module is used to obtain the attention weights from the input feature maps at different scales so that the module can combine contextual information at different scales and generate better pixel-level attention feature maps. In addition, the interaction of attention information is realized and the cross-dimensional vectors are combined without destroying the original channel attention vector. Thus, the entire multi-scale channel attention vector is obtained using concatenation as follows:

$$Z = Z_0 \oplus Z_1 \oplus \cdots \oplus Z_{S-1}, \tag{10}$$

where $\oplus$ denotes the concatenation operation, Z_i is the attention weight value obtained from F_i, and Z is a multi-scale attention weight vector.

To enable each channel to adaptively select different spatial scales, the attention vector of each channel is normalized using Z_i as follows:

$$A_i = \text{softmax}(Z_i) = \frac{\exp(Z_i)}{\sum_{i=0}^{S-1} \exp(Z_i)} \tag{11}$$

where the softmax function is used to obtain the normalized multi-scale channel attention weight A_i, which contains all spatial position information and the attention weight in the channel to enable the interaction between local and global channel attentions. Next, the channel attention of the normalized weight feature map is recombined using concatenation so that the attention weights of the entire channel are expressed as follows:

$$A = A_0 \oplus A_1 \oplus \cdots \oplus A_{S-1}, \tag{12}$$

where A refers to the multi-scale channel weight after the attention interaction. Then, the normalized weights of multi-scale channel attention A_i are multiplied by the feature map of the corresponding scale $F_{i,}$ yielding,

$$Y_i = F_i \cdot A_i, \quad i = 0, 1, 2, \ldots, S-1 \tag{13}$$

where $\cdot$ refers to channel multiplication, Y_i is the feature map generated after the multi-scale attention weights have been obtained and applied to the channel through multiplication. This channel-wise multiplication retains the expression of the overall feature without destroying the original feature map information. Finally, the obtained feature maps Y_i with the new weights are dimensionally concatenated to obtain the final output result as follows:

$$\text{Out} = \text{Concat}([Y_0, Y_1, \cdots Y_{S-1}]) \tag{14}$$

In summary, using a lightweight structure, the PSA module integrates multi-scale spatial information and cross-channel attention into each segmented feature group so that the local and global channel attentions can enable information to better interact and output feature maps with multi-scale, global, and long-term information. In this way, the overall segmentation performance of the model is improved at the cost of only a small increasing in the amount of computation.

3 Experiments and discussion

To verify the effectiveness and segmentation performance of the proposed BHP-UNet model, an experiment was carried out to compare the BHP-UNet to the classic and mainstream segmentation network models FCN, UNet, and DeepLabv3+ based on our side-scan sonar seabed underwater target dataset (shipwreck dataset is used in this experiment). In addition, the segmentation performance was evaluated using seabed shipwreck data measured in a certain sea area of Zhoushan, China. Next, an ablation experiment was performed to evaluate the effectiveness of the BHD and PSA modules in BHP-Unet. Finally, a simulation experiment that simulated complex seabed situations was conducted to further evaluate the segmentation performance of the BHP-UNet model.

3.1 Dataset and experiment configuration

The experimental data used in this paper are composed of data measured in a specific sea area of Zhoushan along with data provided by various international and Chinese marine-related institutions as well as Chinese manufacturers. The Zhoushan data were obtained using an iSide1400 side-scan sonar system from July to August 2022, and part of the images of the field experiment are shown in Fig. 6. One seabed shipwreck target was found in the experiment, and a total of 5 side-scan sonar images of the target from different directions and depths were obtained.

The data provided by marine-related departments and Chinese manufacturers were obtained from measurements in the East China Sea, South China Sea, Huangbohai Sea, and inland lakes, using common side-scan sonar devices including the Klein3000, EdgeTech4200, Yellowfin, and Hydro series. To further enrich the side-scan sonar shipwreck database, a web-crawler program was used to collect data from the Internet, and a total of 1,200 images were obtained. The data provided by marine-related institutions and manufacturers were used for model training and validation. The Zhoushan data were used as the evaluation dataset after the model was trained to evaluate its segmentation performance.

To better analyze the characteristics of the targets in the dataset, the distribution of the shipwreck targets and their length-to-width ratio in the images were investigated. The results are shown in Fig. 7, the depth of color in the visualization corresponds to the quantity of images, with darker shades indicating a higher number of images.

As Fig. 7 shows, the shipwreck targets are mainly concentrated in the center of the image, and most are small targets, including those that are densely arranged and overlapping complex targets. To further enrich the sample data, compensate for the size and distribution limitations of the shipwreck target images, and improve model training, the dataset was first normalized by changing the size of each image to 300×300 pixels, and data augmentation operations like mosaicking, image rotation, multi-scale cropping and magnification, image translation, image flipping, and noise enhancement were carried out. As a result, the number of data images was increased from 1200 to 3000. The mosaic data augmentation method increases the position

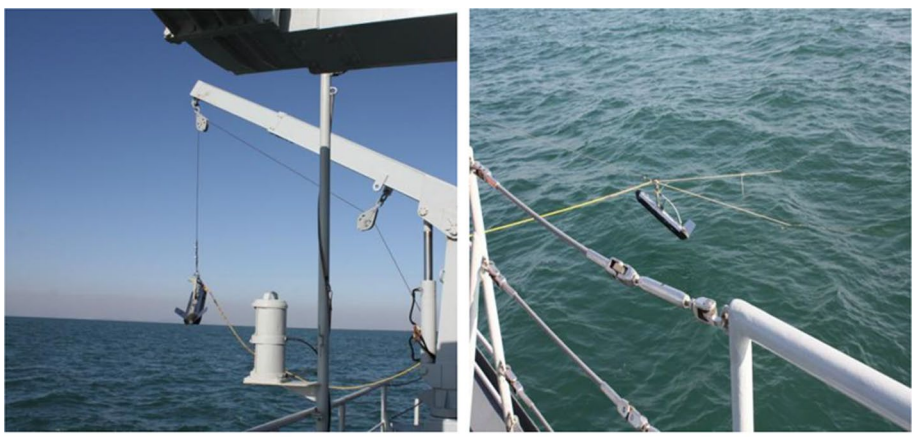

Fig. 6 Images of the field experiment. **a** is side-scan sonar modulation; **b** is side-scan sonar recovery

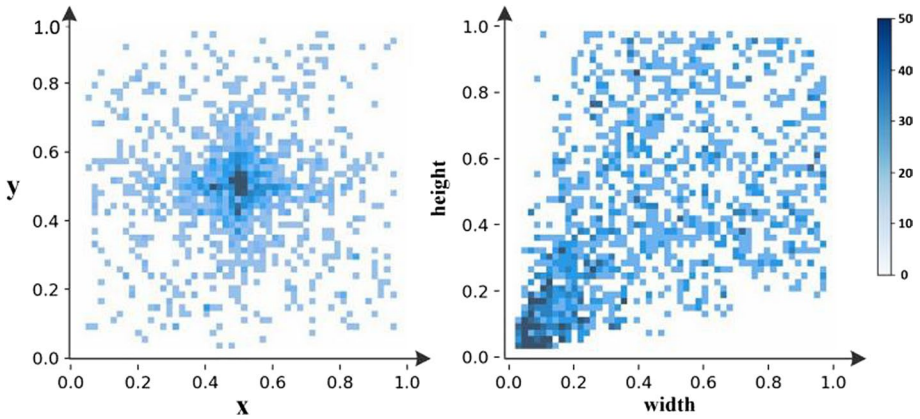

Fig. 7 Distribution and sizes of the shipwreck targets

distribution of the targets and enlarges small targets to a certain extent, thereby improving the generalization ability of the model while improving the efficiency of model training.

The methods in this experiment were implemented in Python and based on the PyTorch framework. The experiments were performed on a computer running the Windows 10 operating system and equipped with an Intel(R) Core(TM) i9-10900X@3.70 GHz CPU, and two NVIDIA GeForce RTX 3090 s GPUs with 48 GB parallel memory.

To improve the training efficiency while ensuring training performance, the training and test sets in the experiment were divided using a ratio of 8:2, and tenfold cross-validation was used for model training. In tenfold cross-validation, the dataset is divided into ten equal subsets, where nine subsets are used for training the model and the remaining subset is used for evaluating model performance. By employing tenfold cross-validation, different combinations of training and validation sets are considered, reducing bias introduced by the dataset selection and obtaining more stable and reliable model performance evaluation. Simultaneously, through grid search, various combinations of hyperparameters are explored, and the best combination is selected based on the evaluation results from the validation set, this allows for further improving its performance and generalization capability. The initial learning rate was set to 0.0001, and a warmup period of five epochs was carried out before the start of training to stabilize the model training and accelerate the convergence of the model. Meanwhile, Adam optimization was used for step annealing to adaptively adjust the learning rate. The number of training steps for each batch was set to 1000 epochs, the batch size was set to 32 according to the computer configuration, which resulted in a total of 75k steps to completely train the model.

3.2 Model training and performance evaluation

This experiment compared the side-scan sonar seabed shipwrecks segmentation performance of four models, FCN, UNet, DeepLabv3+, and BHP-UNet. First, the four models were trained using this experimental dataset, as shown in Fig. 8. Second, the evaluation set was used to evaluate the performances of the four models after training.

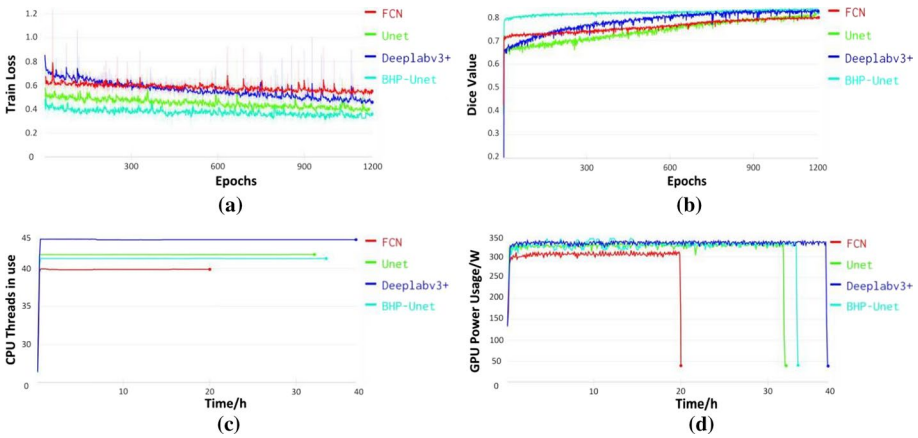

Fig. 8 Comparison of the training processes of the four models. **a** Description of training loss; **b** description of dice scores of the four models on the validation dataset evaluated using tenfold cross-validation. **c** Description of number of CPU threads in use; **d** description of GPU power usage

The objective of the BHP-UNet model is to minimize the discrepancy between the predicted results and the true results in order to optimize the model's parameters. Through the backpropagation algorithm, the model's parameters are updated based on the gradient information of the loss function, allowing the model to more accurately predict binary classification labels. The BHP-UNet model is trained and tested using the binary cross-entropy loss function, which is defined by the following formula:

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log\left(1-\hat{y}_i\right)\right) \tag{15}$$

where y represents the true binary labels, represents the model's predicted outputs, and N is the number of samples.

The binary cross-entropy loss function quantifies the performance of the model by comparing the differences between the predicted values $\hat{y}_i$ and the true values y_i. The first term $-y_i \log(\hat{y}_i)$ of the loss function is active when y_i is 1, while the second term $-(1-y_i)\log\left(1-\hat{y}_i\right)$ is active when y_i is 0.

As Fig. 8a shows, the training loss values of the four models decreased as the number of training steps increased, finally stabilizing and converging. Of the four models, the FCN model had the highest loss value. The DeepLabv3+ model had the largest initial loss value, a slower convergence speed, and longer training time because it has the most model parameters and the most complex structure. In addition, its decrease in amplitude also increased as the number of training steps increased, but many large fluctuations occurred after 700 epochs, indicating a certain instability. The BHP-UNet model had the lowest training loss value, which was about 0.1 lower than that of the conventional UNet model, and the oscillation amplitude was the smallest during the training process. BHP-UNet tended to fit the data after 600 epochs, and the training efficiency was high.

In this experiment, the Dice score was used as an evaluation metric to indicate model segmentation performance. It is calculated as follows:

$$\text{Dice} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{16}$$

where $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$.

Precision refers to the probability that a pixel that is identified as a positive sample is actually a positive sample, and it is used to measure the accuracy of the results. Recall refers to the probability that the positive samples are identified as positive samples and it is used to measure the completeness of the results. TP (true positives) refers to correctly identified positive samples, FP (false positives) refers to incorrectly identified positive samples, TN (true negatives) refers to correctly identified negative samples, and FN (false negatives) refers to incorrectly identified negative samples. $|X|$ denotes ground truth, indicating the real pixels of the target and $|Y|$ denotes the predicted mask, indicating the segmented pixels predicted by the model.

As Fig. 8b shows, the BHP-UNet model obtained the highest Dice score and tended to be stable following 10k training iterations. Moreover, the model converged with a very high training efficiency. The Dice score of the DeepLabv3+ model eventually reached 0.8134, which is only slightly lower than the Dice score of the BHP-UNet model (0.8205). However, the model had a lower training efficiency, with a certain oscillation, and it did not converge.

According to Fig. 8c and d, the training time of the BHP-UNet model was only about 10 min longer than that of the UNet model when the BHD and PSA modules were combined. However, it obtained almost the same results as UNet with respect to the number of CPU threads in use and GPU power usage. By contrast, the DeepLabv3+ model had the longest training time and consumed the most hardware resources because of its more complex model structure.

In summary, compared with the conventional UNet and FCN models, the BHP-UNet model obtains substantially improved performance in terms of the model training loss value, fitting efficiency, and segmentation performance while sacrificing a certain amount of training time and requiring more hardware resources. Compared with the most complex DeepLabv3+ model, the segmentation performance of the BHP-UNet model is better because training time, efficiency, and hardware resources are ensured.

To evaluate the segmentation performance of the model after training, FCN, UNet, DeepLabv3+, and the BHP-UNet model were evaluated using the test set. The accuracy of model segmentation was evaluated using the Dice score, IoU, IoU is calculated as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} = \frac{X \cap Y}{X \cup Y} \tag{17}$$

The efficiency of the model segmentation was evaluated using frames per second (FPS), which represents the number of 300×300 resolution images that can be segmented per second using two NVIDIA GeForce RTX 3090 GPUs. Using the weights of the generated model (Weights) as the basis for evaluating future model development, the specific segmentation and quantification results of the four models are presented in Table 1.

Table 1 Comparison of the segmentation results of the four models in the test set

Method	Dice/%	IoU/%	FPS	Weights/MB
FCN	0.5214	0.5897	**302**	**46.7**
UNet	0.6926	0.7147	126	65.9
Deeplabv3+	0.7680	0.7520	88	101.6
BHP-UNet	**0.7831**	**0.7771**	121	73.2

Bold values indicate the best performance values on Dice, IoU, FPS, and Weights

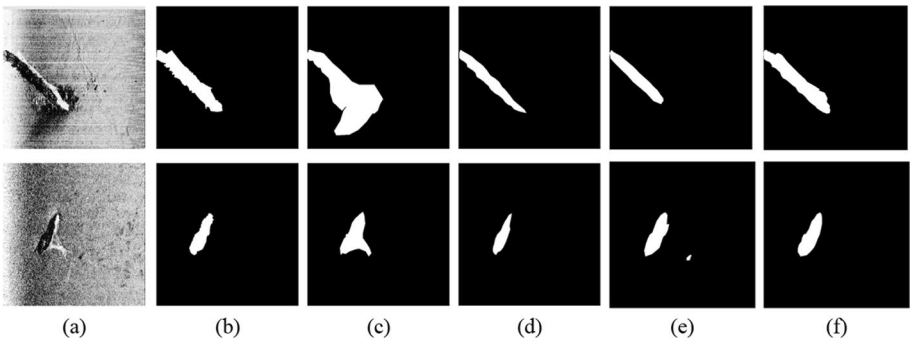

Fig. 9 Comparison of the detection results of the four models. The upper image in **a** shows a target image obtained from the port side, with the direction of 150° east of north and 15 m to the bottom, and the lower image in **a** is a target image obtained from the starboard side, with the direction of 30° west of north and 30 m to the bottom; **b** is the annotation map; the segmentation results of the four models FCN, UNet, DeepLabv3+ and BHP-UNet are shown from (**c**) to (**f**)

Table 1 reveals that the BHP-UNet model had a higher Dice score and IoU result than the other three models. The Dice score was 78.31%, which is an increase of 26.17%, 9.05%, and 1.51%, respectively, in the Dice scores of the other three models (FCN, UNet, and DeepLabv3+). In addition, its IoU was 77.71%, which is an increase in IoU of 18.74%, 6.24%, and 2.51%, respectively, indicating that the BHP-UNet model has the best segmentation performance. The DeepLabv3+ model has the most complex structure and highest number of parameters, but its segmentation performance is not as good as that of the BHP-UNet model. Moreover, its Weights result is much higher, and its FPS is much lower than those of the other three models. These results are not conducive to lightweight engineering applications. The BHP-UNet model adds the BHD and PSA modules, which increases the model complexity, and thereby is bound to cause the model to underperform the FCN and UNet models in terms of FPS and Weights, but the large improvement in segmentation performance at the expense of a small amount of segmentation speed and an increase in the number of model weights is very cost effective. Furthermore, a small decrease in FPS and a small increase in the number of weights with respect to the UNet model will not have a substantial impact on the lightweight engineering deployment of the model.

To further evaluate the segmentation performances of the four models, the Zhoushan evaluation set was used, and the side-scan shipwreck target images from different scanning directions and distances to the bottom were selected.

Figure 9c shows that the FCN model could only obtain the basic position of the target from the segmentation, but it incorrectly segmented the shadow area around the shipwreck target, with a high missed detection rate. Figure 9d reveals that the UNet model could obtain the basic shipwreck target by segmentation, but its results are inferior to those of the BHP-UNet model in terms of segmentation and position accuracy. As Fig. 9e shows, the segmentation effect of the DeepLabv3+ model is significantly better than those of UNet and FCN, but there is still the problem of missed detections. In general, the overall segmentation results of the BHP-UNet model are the closest to the annotation map and are the best. Although the outline details need to be improved when compared with the annotation map, it obtains the most detailed segmentation results, and the segmentation area is the most accurate when compared with the areas obtained by the FCN, UNet, and DeepLabv3+ models.

In summary, the BHP-UNet model obtains the best segmentation for the side-scan sonar seabed shipwreck target among the four models after it was integrated with the BHD and PSA modules, although there is a small increase in training time and number of model weights.

3.3 Ablation experiment and evaluation

To further analyze the reasons for the performance improvements of the BHP-UNet model and to verify the effectiveness of the BHD and PSA modules, an ablation experiment was designed, the hyperparameters employed in these sections are the same as those mentioned in Sect. 2.1. As above, the Dice value and IoU were used as the evaluation metrics, and the control variable method was employed to compare and analyze the effect of each module on its segmentation performance. The experimental results are presented in Table 2. In the experiments, Group 1 represents the model without using BHD module and PSA module. Group 2 represents the model using only the BHD module. Group 3 represents the model using only the PSA module. Group 4 represents the model using both the BHD module and PSA module, which is the proposed model in this paper.

A comparison of Groups 1 and 2 reveals that the integration of the BHD module improves the Dice score and IoU. Of the two, the Dice score was improved more than the IoU, with an increase of 5.38%, which proves that the integration of BHD module enables the model to combine multi-scale features while expanding the receptive field, highlighting the feature areas with a high learning response and learning more complementary feature areas, thereby improving the segmentation accuracy of the model. By comparing Groups 1 and 3, we find that the addition of PSA module has also improved the Dice score and IoU of the model. Of the two, the IoU was increased by 4.51%, which

Table 2 Comparison of the segmentation performance of different strategies

Group	BHD module	PSA module	Dice/%	IoU/%
1	–	–	0.6926	0.7147
2	√	–	0.7464	0.7331
3	–	√	0.7255	0.7598
4	√	√	0.7831	0.7771

proves that the addition of PSA module has enabled better information interaction to be formed between the local and global channel attentions, and multi-scale, global, and long-term connections are established, thereby enabling accurate positioning output to be achieved while the features are well captured. A comparison of Groups 1 and 4 reveals that the combination of both modules increased the Dice score by 9.05% and the IoU by 6.24%. The comparison of the results of Group 4 with those of Groups 2 and 3 shows that the combination of both modules is better than the use of a single module. These results show that the BHD module more obviously improved performance in terms of segmentation accuracy, whereas the PSA module more obviously improved performance in terms of the segmentation and positioning precision.

To further evaluate the underwater target segmentation performance of the model under complex conditions and examine the role and contribution of each module in the model, experiments were conducted using underwater target images exhibiting characteristics such as dense arrangement and overlapping.

To further evaluate the underwater target segmentation performance of the model in complex scenarios and the roles and contributions of each module in the model, experiments were conducted using underwater target images with different complexities, such as densely arranged and overlapping targets. Figure 10 presents a comparison of the partial shipwreck target segmentation results under different strategies. From left to right, the images show the original image, annotated image, results of Unet, results of using only the BHD module, results of using only the PSA module, and results of the BHP-Unet model segmentation.

The three groups of side-scan sonar underwater target images in Fig. 10 represent different complex scenarios, including densely packed, overlapping, and cluttered debris situations. Overall, the Unet model achieves a basic level of segmentation for the underwater targets, but there is room for improvement in capturing fine details, particularly in distinguishing densely packed and overlapping targets. The model incorporating the

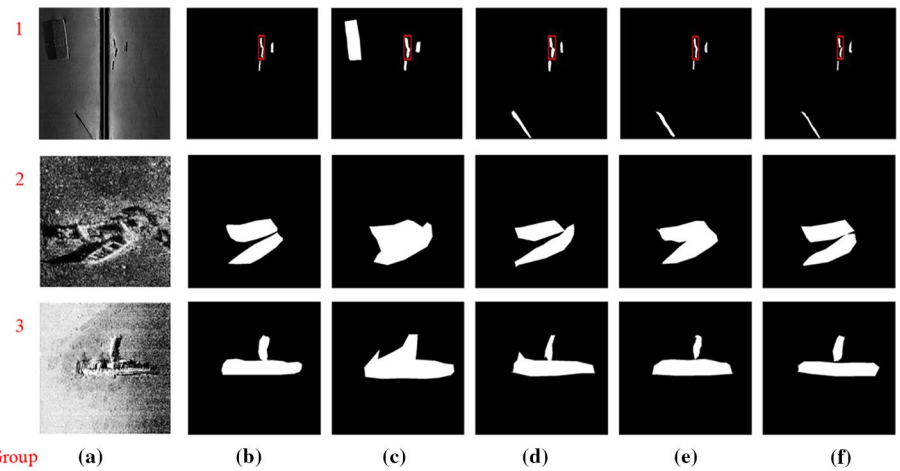

Fig. 10 Compares the segmentation results of some shipwreck targets obtained by the different versions of the model. **a**, **b** is the annotation map; **c** is the segmentation results of UNet; **d** is the segmentation results of model with the BHD module only; **e** is the segmentation results of with the PSA module only, and **f** is the segmentation results of the BHP-UNet model

BHD module shows improved localization accuracy compared to the Unet model. The addition of the PSA module significantly enhances the segmentation accuracy. In comparison, the proposed model, the BHP-Unet model, which combines the BHD and PSA modules, outperforms other models in terms of both accuracy and localization, demonstrating superior segmentation performance.

Taking Group 1 as example, as Fig. 10c shows, the UNet model completed the basic segmentation task but failed to distinguish the two shipwreck targets within the red box, which were densely arranged in a top-to-bottom direction, and mistakenly segmented the two targets as one target. It also made a false detection by mistaking the rectangular stone pier on the left as a shipwreck target. According to Fig. 10d, U-Net with the BHD module successfully distinguished the two shipwreck targets within the red box. It has a higher segmentation accuracy than the UNet model because it did not falsely detect the rectangular stone pier on the left, but it misdetected the submarine pipeline as a shipwreck target. Figure 10e reveals that UNet with the PSA module mistakenly segmented the two targets within the red box as one target and falsely detected the submarine pipeline, but the areas segmented by it are more in line with the actual target areas, and the positioning precision is higher than that of the UNet model. The BHP-UNet model, which incorporates both the BHD and PSA modules, inherits the advantages of both modules, although it also falsely detected the submarine pipeline. It distinguished the densely arranged shipwreck targets well and obtained a more accurate positioning precision, improving both the segmentation accuracy and segmentation positioning precision when compared with the cases in shown in Fig. 10c, d, and e.

3.4 Simulation experiment and evaluation

Because of the existence of various environmental noises in seawater and the time-varying and spatial distortion of underwater acoustic signals, different sea conditions and marine environments will cause different degrees of interference in sonar images. To further evaluate the segmentation performance of the model in a more complex marine environment, an experiment that simulated complex sea conditions was designed.

Because speckle noise is the main factor affecting the quality of side-scan sonar images [36], Rayleigh noise with an expected value of zero and a standard deviation of 60 was added to the Zhoushan seabed shipwreck images in this experiment. The upper image in Fig. 11a is the image obtained from the starboard side, with a direction of due north, 10 m to the bottom, and the lower image is an image obtained from the starboard side, with a direction of 30° west of north, 40 m to the bottom. The segmentation results of the three models are shown in Fig. 11c and d. From left to right, they are the UNet, DeepLabv3+, and BHP-UNet models. FCN was not included in the simulation experiment because of its poor segmentation results in the previous experiment.

According to Fig. 11, using the noisy side-scan sonar shipwreck images, all the three models could obtain the target by segmentation without missed detections. However, both the accuracy and positioning precision need to be improved. It can be seen from Fig. 11c that the segmentation results of the UNet model are significantly inferior to those of the other two models, and there is a false detection in the image segmentation results in the lower image. Figure 11d shows that the segmentation effect of the DeepLabv3+ model is significantly better than that of UNet, but it segmented a single

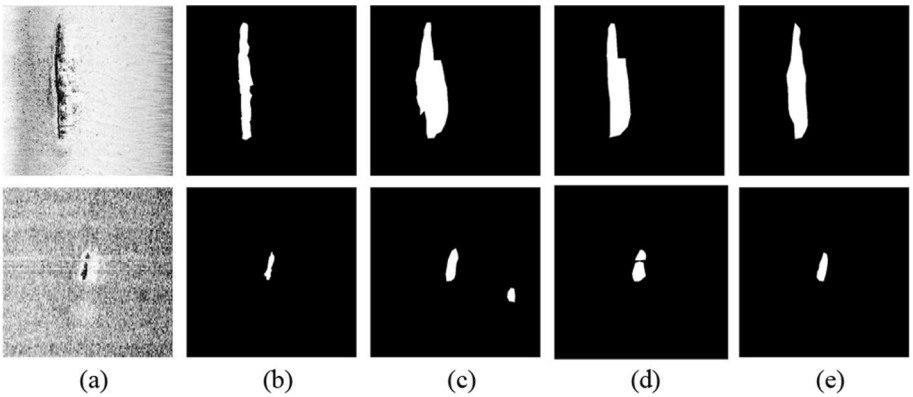

Fig. 11 Comparison of the noisy target segmentation results of three models. **a** is the noisy map; **b** is the annotation map; **c** is the segmentation results of UNet; **d** is the segmentation results of DeepLabv3+; **e** is the segmentation results of BHP-UNet

shipwreck target into two targets in the lower image. Figure 11e reveals that although the segmentation map of BHP-UNet still has a certain gap with respect to the annotation map, the segmentation results are significantly better than those of the UNet and DeepLabv3+ models, reflecting to some extent that BHP-Unet can better adapt to the complex environment of the ocean and has better segmentation performance and generalization ability, with stronger practicability and guiding significance.

4 Conclusion

In order to address the challenges of detecting densely arranged and overlapping underwater seabed targets with high false alarm and miss rates, this study presents a novel underwater target segmentation model based on the BHP-Unet algorithm using side-scan sonar data. The main focus of this model is to enhance the segmentation performance of underwater targets in complex scenarios. The key innovations of our proposed approach are as follows:

- *BHD module* We designed a BHD module that enhances the receptive field while performing multi-scale feature fusion. This module highlights the learning of highly responsive feature regions and captures more complementary feature regions, thereby significantly improving the accuracy of target segmentation.
- *PSA module* We introduced a PSA module, which facilitates better information interaction between local and global channel attention. This module establishes multi-scale, global, and long-term connections, enabling accurate localization outputs while effectively capturing diverse features.
- *BHP-Unet model* By integrating the BHD module, PSA module, and Unet model with side-scan sonar image features, we propose the BHP-Unet model. This innovative fusion approach enhances the segmentation performance of underwater targets in complex scenarios, providing more accurate and reliable results.

To validate the effectiveness and segmentation performance of the BHP-Unet model, this study conducted three sets of experiments:

Firstly, we compared the BHP-Unet model with FCN, Unet, and Deeplabv3+ models on a self-made dataset of side-scan sonar shipwrecks. The experimental results demonstrate that the BHP-Unet model achieved the highest Dice value and IoU on the test set, with a Dice value of 78.31% and an IoU of 77.71%. Moreover, it exhibited superior segmentation performance on side-scan sonar shipwreck images obtained from the Zhoushan sea area.

Secondly, we selected complex side-scan sonar underwater target images characterized by dense arrangement and overlapping and conducted ablation experiments to validate the effectiveness of the BHD module and PSA module. The results showed that the incorporation of the BHD module increased the Dice value by 5.38%, while the inclusion of the PSA Module improved the IoU by 4.51%. The combination of these two modules resulted in an overall improvement of 9.05% in Dice value and 6.24% in IoU, enabling high-performance segmentation in scenarios with densely arranged and overlapping multiple targets.

Lastly, we designed simulation experiments by adding noise to side-scan sonar shipwreck images collected from the Zhoushan sea area to mimic complex underwater conditions. The results demonstrated that the BHP-Unet model exhibited better adaptability to the complex marine environment, showcasing superior segmentation performance and generalization capabilities.

In conclusion, the BHP-Unet model, despite sacrificing a certain degree of training efficiency and model weights, achieved remarkable segmentation performance, addressing the segmentation challenges of densely arranged and overlapping shipwreck targets to a certain extent. However, to further enhance the segmentation performance for small-sized targets in complex marine conditions, techniques such as sonar data augmentation, few-shot learning, and even zero-shot learning methods should be considered. Additionally, to meet the requirements of future lightweight engineering deployment, further research should be conducted on optimizing the model structure and segmentation efficiency.

Abbreviations

BHP-UNet	Blended hybrid dilated convolution and pyramid split attention UNet
BHD	Blended hybrid dilated convolution
PSA	Pyramid split attention
VGG	Visual geometry group
YOLO	You only look once
DETR	Detection transformer
SVM	Support vector machine
FCN	Fully convolutional networks
Faster-RCNN	Fast region-based convolutional network
Deeplab	Deep labeling for semantic image segmentation
ASPP	Atrous spatial pyramid pooling
Upsampling	Upconv
Concatenation	Concat
FP	False positive
TP	True positives
FN	False negatives
TN	True negatives
SE	Squeeze-and-excitation
SPC	Split-and-concatenate
CBAM	Convolution block attention module
FPS	Frame per second

Acknowledgements
We would like to thank the editor and the reviewers for their valuable comments and suggestions that greatly improve the quality of this paper.

Author contributions
YT devised the proposed framework and performed the simulations; LW and HL helped revise the manuscript grammar check; SB helped formulate the problem optimization; YT helped design the networks; LW helped conduct the experiments, and YT helped explain the experiments and discussion outcomes. All authors read and approved the final manuscript.

Funding
This work was funded by National Science Fund of China for Distinguished Young Scholars under Grant 42122025; National Natural Science Foundation of China under Grant 41974005, Grant 42074074 and Grant 41971416.

Availability of data and materials
The data that support the findings of this work are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 March 2023 Accepted: 22 June 2023
Published online: 30 June 2023

References

1. P. Blondel, The Handbook of Sidescan Sonar, (2009), pp. 147–183. doi:https://doi.org/10.1007/978-3-540-49886-5_7
2. M. Zimmermann, H. Deilami, G. Schuster, Automatic Interpretation of Side Scan Sonar Images Using Textural Features and a Neural Network. Oceans 2000 MTS/IEEE Conference and Exhibition. IEEE, (2000).
3. A. Ku, X. Zhou, C. Peng, Research status and development of side-scan sonar detection technology. Hydrogr. Surv. Charting **38**(01), 50–54 (2018)
4. S.G. Johnson, M.A. Deaett, The application of automated recognition techniques to side-scan sonar imagery. IEEE J. Ocean. Eng. **19**(1), 138–144 (1994)
5. H. Li, J. Gao, W. Du, Object representation for multi-beam sonar image using local higher-order statistics. EURASIP J. Adv Signal Process. (2017). https://doi.org/10.1186/s13634-016-0439-7
6. Y. Sun, X. Liu, F. Zhang, Z. Qiu, Side-scan sonar sounding system with shallow high resolution and its marine applications. Ocean Eng. **27**(4), 96–102 (2009)
7. M.R. Arshad, Recent advancement in sensor technology for underwater applications. Indian J. Mar. Sci. **38**(3), 267–273 (2009)
8. H.J. Flowers, J.E. Hightower, A novel approach to surveying sturgeon using side-scan sonar and occupancy modeling. Mar. Coast. Fish. **5**(1), 211–223 (2013)
9. J. C. Isaacs, Sonar automatic target recognition for underwater UXO remediation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2015), pp. 134–140.
10. X. Wang. Research on Precise Processing of Side Scan Sonar Image and Object Recognition Methods. Wuhan University, 2017.
11. P. P. Zhu, J. Isaacs, B. Fu, S. Ferrari, Deep learning feature extraction for target recognition and classification in underwater sonar images. IEEE 56th Annual Conference on Decision and Control (CDC) (2017), pp. 2724–2731.
12. E. Dura, Y. Zhang, X.J. Liao, G.J. Dobeck, L. Carin, Active learning for detection of mine-like objects in side-scan sonar imagery. IEEE J. Ocean. Eng. **30**(2), 360–371 (2005)
13. Y. Song, B. He, P. Liu, Real-time object detection for AUVs using self-cascaded convolutional neural networks. IEEE J. Ocean. Eng. **46**(1), 56–67 (2021)
14. J.M. Topple, J.A. Fawcett, MiNet: efficient deep learning automatic target recognition for small autonomous vehicles. IEEE Geosci. Remote Sens. Lett. (2020). https://doi.org/10.1109/LGRS.2020.2993652
15. P. Feldens, A. Darr, A. Feldens, F. Tauber, Detection of boulders in side scan sonar mosaics by a neural network. Geosciences **04**, 159 (2019)
16. L. Zheng, K. Tian, Detection of small objects in sidescan sonar images based on POHMT and Tsallis entropy. Signal Process. Off. Publ. Eur. Assoc. Signal Process. (EURASIP) **142**, 168–177 (2018)
17. Y. Steiniger, D. Kraus, T. Meisen, Survey on deep learning based computer vision for sonar imagery. Eng. Appl. Artif. Intell. Int. J. Intell. Real-Time Autom. (2022). https://doi.org/10.1016/j.engappai.2022.105157
18. Y. Tang, S. Jin, G. Bian, The transfer learning with convolutional neural network method of side-scan sonar to identify wreck images. Acta Geod. Cartogr. Sin. **50**(2), 260–269 (2021)

19. Y. Tang, S. Jin, G. Bian, Y. Zhang, F. Li, Wreckage Target Recognition in Side-scan Sonar Images Based on an Improved Faster R-CNN Model. International Conference on Big Data & Artificial Intelligence & Software Engineering (2020), pp. 348–354.
20. Y. Tang, S. Jin, G. Bian, Y. Zhang, Shipwreck target recognition in side-scan sonar images by improved YOLOv3 model based on transfer learning. IEEE Access **08**, 173450–173460 (2020)
21. Y. Tang, B. Shao, Z. Guo, M. Liu, W. Zhang, Improved YOLOv5 method for detecting shipwreck target with side-scan sonar. Geomat. Inf. Sci. Wuhan Uni. (2021). https://doi.org/10.13203/j.whugis20210353
22. Y. Tang, H. Li, W. Zhang, S. Bian, G. Zhai, A lightweight DETR-YOLO method for detecting shipwreck targets with side-scan sonar. J. Syst. Eng. Electron. **44**(08), 2427–2436 (2022)
23. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (2015), pp. 3431–3440.
24. O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation. Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (2015), pp. 234–241.
25. L. Chen, G. Papandreou, I. Kokkinos, et al., Semantic image segmentation with deep convolutional nets and fully connected CRFs. Proceedings of International Conference on Learning Representations (2015).
26. L. Chen, G. Papandreou, I. Kokkinos et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
27. L. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation (2017). arXiv: 1706.05587
28. L. Chen, Y. Zhu, G. Papandreou et al., Encoder-decoder with atrous separable convolution for semantic image segmentation (2018). arXiv:1802. 02611
29. F. Yuan, F.Q. Xiao, K.H. Zhang, Y.F. Huang, E. Chen, Noise reduction for sonar images by statistical analysis and fields of experts. J. Vis. Commun. Image Represent. **74**, 102995 (2021)
30. F. Maussang, M. Rombaut, J. Chanussot, Fusion of local statistical parameters for buried underwater mine detection in sonar imaging. EURASIP J. Adv. Signal Process. **2008**, 876092 (2008). https://doi.org/10.1155/2008/876092
31. S. Leier, A.M. Zoubir, Aperture undersampling using compressive sensing for synthetic aperture stripmap imaging. EURASIP J. Adv. Signal Process. **2014**, 156 (2014). https://doi.org/10.1186/1687-6180-2014-156
32. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions (2016). arXiv:1511.07122
33. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks (2018). arXiv:1709.01507
34. S. Woo, J. Park, J. Lee, I. S. Kweon, CBAM: convolutional block attention module (2018). arXiv:1807.06521
35. I.C. Duta, L. Liu, F. Zhu, L. Shao, Pyramidal convolution: rethinking convolutional neural networks for visual recognition (2020). arXiv:2006.11538
36. L. Hellequin, J.M. Boucher, X. Lurton, Processing of high-frequency multibeam echo sounder data for seafloor characterization. IEEE J. Ocean. Eng. **28**(1), 78–89 (2003)

Publisher's Note

Yulin Tang received the B.S. degree in Dalian Naval Academy, and he is currently studying for a Ph.D. degree at the Naval University of Engineering, China. His main research direction is side-scan sonar image processing, underwater target detection, AUV system, and computer vision. E-mail: yltanghg@163.com.

Liming Wang is a professor at the Naval University of Engineering, China, doctoral supervisor, chief scientist of the National Key R&D Program, and expert in the fields of intelligent equipment and unmanned platforms; mainly engaged in the fields of ship intelligent monitoring technology, intelligent detection robots, etc.

Houpu Li is a professor at the Naval University of Engineering, China. He got founded by National Science Foundation for outstanding Young Scholars in 2021. He specializes in the mathematical analysis of geodesy.

Shaofeng Bian is a professor at the Naval University of Engineering, China. He was awarded an Alexander von Humboldt Research Fellowship in 1996. He got founded by National Science Foundation for Distinguished Young Scholars in 2001.