Semantic segmentation method of underwater images based on encoder-decoder architecture

Jinkang Wang, Xiaohui He*, Faming Shao, Guanlin Lu, Ruizhe Hu, Qunyan Jiang

Department of Mechanical Engineering, College of Field Engineering and Army Engineering University, PLA, Nanjing, China

* gcbbxh@aeu.edu.cn

Abstract

With the exploration and development of marine resources, deep learning is more and more widely used in underwater image processing. However, the quality of the original underwater images is so low that traditional semantic segmentation methods obtain poor segmentation results, such as blurred target edges, insufficient segmentation accuracy, and poor regional boundary segmentation effects. To solve these problems, this paper proposes a semantic segmentation method for underwater images. Firstly, the image enhancement based on multi-spatial transformation is performed to improve the quality of the original images, which is not common in other advanced semantic segmentation methods. Then, the densely connected hybrid atrous convolution effectively expands the receptive field and slows down the speed of resolution reduction. Next, the cascaded atrous convolutional spatial pyramid pooling module integrates boundary features of different scales to enrich target details. Finally, the context information aggregation decoder fuses the features of the shallow network and the deep network to extract rich contextual information, which greatly reduces information loss. The proposed method was evaluated on RUIE, HabCam UID, and UIEBD. Compared with the state-of-the-art semantic segmentation algorithms, the proposed method has advantages in segmentation integrity, location accuracy, boundary clarity, and detail in subjective perception. On the objective data, the proposed method achieves the highest MIOU of 68.3 and OA of 79.4, and it has a low resource consumption. Besides, the ablation experiment also verifies the effectiveness of our method.

1. Introduction

With the rapid social development, production resources are increasingly scarce, and the development and utilization of marine resources become important for human society. Countries around the world are paying more and more attention to the exploitation of ocean resources. More and more researchers begin to pay attention to the field of underwater vision. The current computer vision studies on underwater images mainly include underwater image enhancement [1–3] and underwater object detection [4, 5]. The former is to improve the
image quality, and the latter is to perform object recognition and localization. This paper focuses on a new research field of underwater vision, i.e., underwater image semantic segmentation, which aims to classify objects at the pixel level, provide fine target contours for underwater images, and enhance the discrimination between background and objects.

As an important branch of image processing, semantic segmentation [6] divide an image into disjoint and meaningful sub-regions, where the pixels in the same region have a certain correlation and the pixels in different regions have certain differences. That is, semantic segmentation is the process of assigning the same labels to the pixels with the same properties.

Due to the complex underwater environment and the scattering and absorption of light by the water medium, underwater images have poor quality. There are problems such as color distortion, low contrast, noise, and uneven illumination, as shown in Fig 1. This poses a great challenge to the integrity and accuracy of underwater image segmentation. In traditional semantic segmentation, some key information is lost because the resolution of the feature graph is reduced by deep convolution networks. The objects of different scales require to combine global and local information, which makes feature extraction difficult. The spatial invariance of CNN leads to a decrease in positioning accuracy and the blurring of the boundary between the object and the background [7].

For low-quality underwater original images, the traditional semantic segmentation methods have poor segmentation results because they cannot deal with complex underwater environment information [8, 9]. To solve the above problems and improve the semantic segmentation effect of underwater images, this paper proposes a new encoder-decoder network structure for the semantic segmentation of underwater images. In this network, the backbone network Resnet obtains the boundary feature map of the images; the densely connected hybrid atrous convolution expands the receptive field; the cascade atrous convolution space pyramid pooling module considers object regions of different scales and optimizes deep semantic features; the context information aggregation mechanism fuses the feature information of the deep network and the shallow network to recover the detailed information of the objects. Experimental results show that compared with the advanced semantic segmentation methods, the proposed method has significant advantages. Also, the effectiveness of each module in the proposed method is verified by the ablation experiment.

The contributions of this paper are summarized as follows:

1. According to the quality degradation characteristics of underwater images, this paper proposes to enhance the images before semantic segmentation, which is original in existing underwater image semantic segmentation methods.

2. The densely connected hybrid atrous convolution can effectively solve the "gridding issue" caused by stacking multi-layer atrous convolution and avoid convolution degradation caused by excessive atrous convolution rate.
3. The cascaded atrous convolution spatial pyramid pooling module can consider the object areas of different sizes, refine the feature maps, and enrich the detailed information of the objects.

4. The context information aggregation mechanism uses a bottom-up method to aggregate the feature information of the deep network and the shallow network to make the object boundary clearer and refine the segmentation results.

This paper is organized as follows. Section 2 briefly reviews the problem of underwater image semantic segmentation and the present research, including the development of semantic segmentation and the latest method of underwater image segmentation. Section 3 introduces the network model proposed in this paper. Section 4 compares our method with other state-of-the-art underwater image segmentation methods and conducts ablation experiments to verify the contribution of each component in the proposed method. Section 5 summarizes the paper and determines the future work.

2. Related work

Semantic segmentation [10], a classic computer vision task, provides pixel-level image understanding in a human-perceptual manner by transforming image pixel information into a mask with highlighted regions of interest. Each pixel in the image is assigned a class ID according to the object it belongs to.

Early segmentation algorithms mainly include grayscale segmentation [11], conditional random field [12], etc. These traditional algorithms have drawbacks of high storage overhead and low computational efficiency, and the segmentation effect is not satisfactory. With the continuous development of deep learning, the image semantic segmentation method based on deep learning emerges. Long et al. designed the first real deep learning semantic segmentation model FCN [13], which is improved based on the VGG-16 network [14] adopts a cross-layer method and considers global semantic information and local location information. However, due to a series of pooling operations, the resolution of the feature map is reduced, and the pixel spatial position information is lost, resulting in a rough segmentation effect. In view of the shortcomings of FCN, the Deeplab series of algorithms [15–17] effectively expands the receptive field and captures image context information through an image pyramid [18], atrous convolution [19], and atrous space pyramid pooling [20], obtaining segmentation results with improved spatial accuracy. However, the dense network structure leads to slow segmentation, and the Deeplab series of algorithms do not perform well on small-sized objects. Dilation10 [21] optimized the convolutional structure by appropriately discarding the pooling layer. It effectively slows down the reduction of the resolution of the feature map and increases the receptive field, but the continuity of the local information of the pixel is affected, leading to its sensitivity to unknown deformation. Segnet [22] adopts an encoder-decoder mechanism and restores the spatial position of pixels through up-pooling. It improves the segmentation resolution, but too many network training parameters lead to high computational costs. RefineNet [23] proposed a multi-path optimization network that can effectively obtain the context information of the image and improve the utilization of local and global features. However, the boundary information of the segmentation object will be partially lost. EncNet [24] introduces a context encoding module, which captures the scene contextual semantics and selectively highlights the feature maps related to the categories, making the segmentation results more refined. DenseASPP [25] adopts dense connection based on ASPP and covers a wide range of semantic information. It achieves a good segmentation effect on high-resolution images, but dense connection also leads to the rapid increase of computation overhead. DFANet [26]
adopts Xception as the backbone structure and incorporates high-level context into the encoder to achieve an effective balance between segmentation speed and accuracy. DANet [27] adopts a dual-attention network with a self-attention mechanism. It enhances the discriminative ability of scene segmentation feature representation and significantly improves segmentation results. Auto-DeepLab [28] extends NAS [29] to semantic segmentation networks and searches for the optimal network architecture automatically, which gradually improves segmentation accuracy. APCNet [30] uses multiple well-designed adaptive context modules to adaptively construct multi-scale context vectors guided by the global image representation. CANet [31] develops a novel dual-branch dense comparison module to effectively utilize multi-level feature representations from CNNs for dense feature comparison. Also, it adds the attention mechanism to fuse information from different supporting examples. MagNet [32] proposes a multi-scale framework, which gradually refines the segmentation output through multiple amplification levels and spreads information from coarse to fine. The problem of local image blurring is solved, and the segmentation performance is greatly improved. SETR (Segmentation TRansformer) [33] regards semantic segmentation as a sequence-to-sequence prediction task and encodes images into a series of patches by deploying a pure transformer. The segmentation accuracy is improved by modeling the global context in each layer of the transformer. RobustNet [34] proposes a new case-selective whitening loss, which separates the domain-specific styles encoded in the high-order statistics of feature representation from the domain-invariant content, and selectively deletes only the style information that leads to domain drift. With this loss function, the robustness of the invisible region segmentation network is improved. To sum up, image semantic segmentation methods based on deep learning show great advantages of autonomous learning and classification of robust features, and they achieve better segmentation accuracy and speed than traditional semantic segmentation methods.

Underwater images are difficult to obtain, and compared with conventional images, underwater images are more difficult to segment due to their low resolution, low contrast, uneven illumination, and color distortion. Aiming at the difficulty of underwater image segmentation, Li et al. [35] designed a new weight function and improved the adaptive GACV image segmentation algorithm, which achieves a better segmentation effect on underwater blurred images. Yan et al. [36] solved the problem of underwater image segmentation based on the WOA algorithm. The method can avoid premature convergence and obtain the global optimal solution, which shows a good segmentation effect and high robustness. Chen et al. [37] improved the K-means algorithm to solve the problem of improper determination of the K value in the grayscale quantization process, which reduces the influence of the initial centroid position of the image and greatly improved the segmentation speed and accuracy. Ma et al. [38] proposed the two-dimensional fuzzy Otsu algorithm, which combines the dual advantages of the classic method and fuzzy theory, and the detection effect is more stable and extensive. Zhu et al. [39] combined the local statistical active contour model with co-saliency detection, and the segmentation efficiency and quality are excellent. Liu et al. [40] introduced the unsupervised color correction method into the Deeplab V3+ encoder structure, which improves the segmentation accuracy of the object boundary.

In general, the semantic segmentation of underwater images based on deep learning methods has greatly improved the segmentation quality of underwater images. This paper selects some representative works and briefly introduces their methods, advantages, and disadvantages in Table 1. However, the simple semantic segmentation of low-quality underwater images cannot achieve good results. Based on the above literature, this paper enhances the quality of underwater images first and then proposes the semantic segmentation method. First, this paper
adopts densely connected hybrid atrous convolution, which can cover a larger receptive field and more effectively maintain the integrity of feature information and reduce information loss than simple multi-layer hole convolution such as Deeplab series algorithms. Secondly, the cascaded atrous convolutional spatial pyramid pooling module refines the feature map and shares the multi-scale target information by fusing the image features under different receptive fields. This is novel for the existing studies. Finally, the context information aggregation network is put forward. By fusing the feature information in the deep network and the shallow network, the details are richer and the edge details are more fully captured.

3. The recommended method’s overview

To solve the difficulties in the semantic segmentation of underwater images, this paper proposes an encoder-decoder architecture for underwater image semantic segmentation based on context information fusion. The flowchart of the algorithm is shown in Fig 2. Firstly, a multi-space transformation underwater image enhancement (UIE) module based on weighted fusion is proposed to enhance the original underwater images. Then, the enhanced images obtain the boundary feature map through the backbone network Resnet [41] with 50 network layers, $1 \times 1$ and $3 \times 3$ filters, and a stride of 2. The feature maps are sent to the densely connected hybrid atrous convolution module (DCHAC) to expand the neuron receptive field and slow down the resolution reduction [42]. The HAC block contains three $3 \times 3$ atrous convolutions, and the atrous rates of the atrous convolutions are 1, 2, and 5, respectively. Next, the cascaded atrous convolutional spatial pyramid pooling module (CASPP) integrates boundary features of different scales to enrich object details, which contains two $1 \times 1$ convolutions and three $3 \times 3$ convolutions.
atrous convolutions with atrous rates of 5, 9, and 13 respectively, and the number of filters is 256. Finally, the context information aggregation decoder (CIAD) is proposed to fuse the features of the shallow network and the deep network, which extracts rich contextual information, and uses $3 \times 3$ convolution to fine-tune the features, which fully grasps the edge details and greatly reduces information loss. Finally, the segmentation effect map is obtained by bilinear interpolation with two times up-sampling. The size of the segmentation map is consistent with that of the original image.

3.1 Multi-space transformation underwater image enhancement module

For underwater images with low resolution, low contrast, uneven illumination, and color distortion, this paper first enhances the quality of the original underwater images. The flowchart of the multi-space transformation underwater image enhancement algorithm is shown in Fig 3.
First, white balance processing is performed to correct the color cast of the image [43]. The purpose of white balance is to eliminate or reduce the color cast caused by the refraction of light. In this process, the gray world method is exploited to remove the blue-green appearance and compensate the red channel of the image to eliminate the red artifact. The red channel compensation formula is as follows:

$$I_{rc}(A) = I_r(A) + (\bar{I}_g - \bar{I}_r) \times (1 - I_r(A)) \times I_g(A)$$  \hspace{1cm} (1)

where, $I_{rc}(A)$ is the red channel value at point A after compensation; $I_r(A)$ and $I_g(A)$ are respectively the values of the red and green channels at point A of the original image; $\bar{I}_g$ and $\bar{I}_r$ are respectively the average values of the green and red channels of the original image.

Then, the following operations are performed on the images after white balance processing:

1. Convert the image from RGB space to LAB space and process the L channel with $\gamma$ correction and then convert the image back to RGB space. $\gamma$ correction can increase the scale of the region of interest in the image. It has a good dynamic compression effect and can adjust the overall brightness of the image by using two variable parameters $\gamma$ and $c$. The expression of $\gamma$ transform is as follows,

$$I_1 = cI_i^\gamma$$  \hspace{1cm} (2)

where, $I_i$ represents the intensity of the input image; $I_1$ represents the intensity of the output image; $\gamma$ and $c$ are used to adjust the shape of the $\gamma$ function.

2. The contrast-limited adaptive histogram equalization algorithm (CLAHE) [44] and bilateral filtering [45] are carried out in the RGB space. White-balanced images still have noise and low contrast. CLAHE can obtain enhanced images with uniform brightness, and bilateral filtering can remove noise and preserve edges to improve image clarity. The output after this operation is denoted as $I_2$.

3. Convert the image from RGB space to HSV space and process the V-channel with single-scale retinex (SSR) [46] and then convert the image back to RGB space. The HSV color space is more in line with the visual characteristics of human beings. Converting the underwater image to HSV space, keeping the hue channel H component and saturation channel S component of the image unchanged, and enhancing the luminance value channel V with SSR combined with guided filtering [47] can help to obtain images suitable for human observation. The output after this operation is denoted as $I_3$.

The outputs $I_1$, $I_2$, and $I_3$ are weighted and fused to obtain the following results:

$$I_0 = aI_1 + bI_2 + cI_3$$  \hspace{1cm} (3)

where $a$, $b$, and $c$ are control parameters, and their sum is 1. Through experiments, it is indicated that when $a = 0.4$, $b = 0.3$, and $c = 0.3$, the image enhancement effect is the best. Fig 4 shows the effect comparison before and after enhancement.

### 3.2 Densely connected hybrid atrous convolution module

Deeplab series networks use multi-layer continuous atrous convolution for feature extraction. It enlarges the receptive field but causes the “gridding issue” [48]. This is because after using multi-layer atrous convolution, the input samples become sparse, which leads to the loss of some local information and the neglect of the overall continuity information of the feature.
map. As shown in Fig 5, the three-layer atrous convolution with a size of 3×3 and a rate of 2 is taken as an example. The continuous atrous convolution covers the incomplete area of the feature map, and the extracted features are lost or even wrong. Thus, it is difficult to achieve a good segmentation effect by using sparse pixel information.

To solve this problem, this paper adopts hybrid atrous convolution with different atrous rates. A small atrous rate is used to extract local information, and a large atrous rate is used to extract long-distance information. This can help to obtain wider regional information and improve the information utilization rate while expanding the receptive field. As shown in Fig 6, the hybrid atrous convolution with atrous rates of 1, 2, and 3 is taken as an example. The coherent receptive field completely covers the whole feature map, which ensures the integrity of information extraction. The hybrid atrous convolution structure is expressed as follows.

\[
M_i = \max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i]
\]

where, \(r_i\) is the atrous rate of layer \(i\), and \(M_i\) is the maximum atrous rate of layer \(i\). For the convolution kernel size of \(k \times k\), \(M_i \leq k\) should hold so that all the receptive fields can be covered.

This paper connects three hybrid atrous convolution blocks through dense connection to encode multi-scale semantic information. Based on this, the final output features cover a large range of semantic information and cover information coding in a very dense way.
3.3 Cascaded atrous convolutional spatial pyramid pooling module

For low-quality underwater images, layer-by-layer convolution gradually reduces the resolution of the feature map, which will lead to the loss of segmentation details and the weakening of the correlation of each pixel. To address this issue, this paper parallels three atrous convolution blocks with atrous rates of 5, 9, and 13 to refine the feature map. The output feature maps under different receptive fields are cascaded and fused to share multi-scale target information, which further improves the relevance and continuity of global pixel information. Finally, the obtained five feature maps are fused in the channel dimension, which involves the object areas of different sizes, enriches the object details, and helps to obtain higher-level image semantic information. The structure of the CASPP module is shown in Fig 7.

3.4 Context information aggregation decoder

The feature maps of different network depths have different characteristics. The shallow feature map is large in scale and has more feature information, which can be used to easily distinguish the object. However, due to the limitation of convolution kernels and computing resources, the shallow network can only extract local features with detailed information, such
as color and texture within a small receptive field. After layer-by-layer convolution and down-
sampling, the resolution and the size of the deep network feature map are reduced, and the
receptive field is enlarged. The deep network feature can extract relatively high-level
semantic information such as location and category in a global scope. However, due to the lack
of geometric space detail, the accuracy of edge segmentation is not enough. The fusion of fea-
ture information in the deep network and the shallow network can greatly reduce the loss of
image feature information, leading to richer detail information richer and clearer edge seg-
mentation results [49, 50].

The structure of the context information aggregation decoder is shown in Fig 8. First, the
feature map F4 output by the CASPP module and the feature map F3 extracted by the hybrid
atrous convolution block 3 are added pixel by pixel. Then, the aggregated feature map is
upsampled two times to increase the resolution of the feature map, and the atrous convolution
with an atrous rate of 2 is used to combine the feature information of adjacent pixels to refine
the upsampling features. Next, the above operations are repeated with the feature map F2
extracted by the hybrid atrous convolution block 2. The obtained feature map is added with F1
extracted by the hybrid atrous convolution block 1, and the result is finally upsampled four
times to aggregate the whole network context. The context information aggregation decoder
gradually merges the deep semantic information with the shallow edge line, shape position,
and other detailed information. This helps to capture clear object boundary information,
refine segmentation results, and effectively improve the accuracy of object segmentation.

4. Experimental results
4.1 Dataset
To complete the research on semantic segmentation of underwater images, this paper com-
bined several open-source underwater image datasets to construct an underwater image seg-
mentation dataset, and the composition of the dataset is shown in Table 2. This paper selected

![Fig 8. The structure of the context information aggregation decoder.](https://doi.org/10.1371/journal.pone.0272666.g008)

| Dataset          | Quantity | Marked quantity | Resolution ratio | Scene distribution          | Source                              |
|------------------|----------|-----------------|------------------|-----------------------------|------------------------------------|
| RUIE [53]        | 4000     | 500             | 400×300          | Scallop, kelp, sea urchin    | Dalian University of Science and Technology |
| HabCam UID [54]  | 10465    | 2300            | 2720×1024        | Fish, scallop               | CVPR AAMVEM studio                 |
| UIEBD [1]        | 950      | 200             | Multi-resolution | Diver, statue, marine life   | City University of Hong Kong        |

https://doi.org/10.1371/journal.pone.0272666.t002
3,000 images for annotation. The “Resize” operation was used to cluster large-resolution images, and pixel interpolation was performed on small-resolution images. Then, the image information was extracted, and the pixels were rearranged to the resolution of 512×512 to achieve a uniform distribution. 3,000 of the images were marked by manual labeling with the LabelMe software. The labeling includes 12 categories: fish, diver, coral, rock, sculpture, octopus, turtle, seaweed, manta ray, starfish, shell, and sea urchin. The labeling format of the COCO dataset [51] was used. To facilitate the research of semantic segmentation, the mask images of semantic segmentation similar to those in the PASCAL VOC 2012 dataset were provided [52]. The ratio of the training set to the testing set is 4:1. The category statistics are shown in Fig 9.

4.2 Experimental setting

Our proposed method was tested in PyTorch software on a computer equipped with Intel Core i7-6700U CPU (4.00 GHz), NVIDIA GeForce RTX 4000, 8 GB DDR3 memory, and running Windows 10 64-bit operating system.

As for training parameter settings, the initial learning rate was set to 0.01, and the model was trained for 30 epochs using the learning rate cosine decay strategy. The network was optimized using SGD with momentum. The momentum parameter was set to 0.9, the weight decay regularizer was set to 0.00005, and the batch size was set to 16.

4.3 Performance evaluation index

This paper used $F_1$, $OA$, $IOU$, and $MIOU$ as comprehensive evaluation indexes to evaluate the segmentation effect of our proposed method. These indexes are widely used in many studies [55, 56]. The ground truth was obtained through manual labeling. $TP$ indicates the number of samples whose prediction results are true and the ground truths are true; $FN$ indicates the number of samples whose prediction results are false and the ground truths are true; $FP$ indicates the number of samples whose prediction results are true and the ground truths are false; $TN$ indicates the number of samples whose prediction results are false and the ground truths are false. $R$ represents the recall rate, and its calculation is shown in Formula (5). $P$ represents the precision, and its calculation is shown in Formula (6). $F_1$ comprehensively evaluates the recall rate and the precision, and its calculation is shown in Formula (7).

$$R = \frac{TP}{TP + FN} \quad (5)$$

$$P = \frac{TP}{TP + FP} \quad (6)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (7)$$
$F1 = \frac{2R \times P}{R + P}$  \hspace{1cm} (7)

$OA$ represents overall accuracy, and it indicates the proportion of correctly classified samples to all samples. The calculation formula is shown as follows.

$$OA = \frac{TP + TN}{TP + FN + FP + TN}$$ \hspace{1cm} (8)

$IOU$ represents the intersection over the union between the predicted value and the real value of each category, and its calculation is shown in Formula (9). $MIOU$ is the average value of $IOU$ in all different categories, and its calculation is shown in Formula (10).

$$IOU = \frac{TP}{FN + FP + TP}$$ \hspace{1cm} (9)

$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP}$$ \hspace{1cm} (10)

where $k$ represents the number of categories, and there are $k+1$ categories if the background is included.

4.4 Analysis of experimental results

4.4.1 Subjective evaluation. To intuitively show the effectiveness of our proposed method, the comparison results between our method and four state-of-the-art underwater image semantic segmentation methods (Deeplab V3+, DFANet, APCNet, and Liu et al. [40]) on the composite dataset are shown in Figs 10 and 11. From left to right are the original image, the enhanced images, and the results generated by Deeplab V3+, DFANet, APCNet, the method proposed by Liu et al., our method, and the ground truth.

Fig 10. Qualitative comparisons on colored cast underwater images. From left to right are original images, the enhanced images, and the results generated by Deeplab V3+, DFANet, APCNet, the method proposed by Liu et al., our method, and the ground truth.

https://doi.org/10.1371/journal.pone.0272666.g010
Due to the color cast, the contrast of the underwater image is reduced. From the experimental results in Fig 10, it can be seen that our method achieves the best segmentation effect on underwater images with a serious color cast. Vertically, Deeplab V3+ misses some image information due to the decrease in contrast, and it cannot fully segment each category. The rock in image (c), the fish on the upper side in image (d), the starfish, and the shell in image (e) are not segmented correctly, and the edges are inaccurately positioned. DFANet performs slightly better than Deeplab V3+, but incorrectly classifies octopuses in image (c) and fails to capture the smaller starfish and shells in image (e). APCNet's segmentation range is too large for already captured object categories, such as the expansion of the diver’s segmentation boundary in image (a). Compared with the previous methods, the algorithm proposed by Liu et al. [40] achieves a better segmentation effect, but it cannot achieve a good segmentation effect for images with low contrast such as image (c). Compared with other algorithms, our method has the advantages of complete object segmentation and accurate positioning. Although there are problems of misjudgment and unclear boundaries, the segmentation results obtained by our method are the closest to the ground truth. Horizontally, for images with blurred boundaries such as images (a), (b), and (d), the segmentation result of the proposed method is clear and closest to the ground truth; For images with low contrast such as image (c) and (e), the effect of image enhancement is obvious. Our method can accurately...
locate some small targets that are easy to miss, while other methods have different degrees of misjudgments and omissions.

Here, a subjective visual comparison was performed on the segmentation effect of the underwater images with higher resolution. Vertically, the segmentation effect of DeepLab V3+ is not ideal. There are different degrees of missed classification, and the classification result is not accurate. For example, the sculpture in image (c) is misclassified as a diver. DFANet achieves a slightly better segmentation effect, but there are also false classifications, e.g., the small sculpture in image (b) and the coral in image (d) are both misclassified as a rock. The segmentation effect of APCNet and Liu et al. [40] is better, but the details still need to be improved. For example, the two small fishes in front of the rock in image (b) cannot be segmented, and the segmentation shapes of the starfish and sea urchins in image (g) are quite different from the ground truth. Compared with the above four methods, our method performs better with clear boundaries and accurate positioning, but there are still some problems, such as inconspicuous details and incomplete object segmentation. Horizontally, although the segmentation effect of our method is the closest to the ground truth, the difference with other methods is not obvious, which is also consistent with our expectations. It can be seen that image enhancement plays an important role in the semantic segmentation of underwater images.

Overall, our method is the best in terms of segmentation completeness, localization accuracy, boundary definition, and detail.

4.4.2 Objective evaluation. The $F_1$, $OA$, $IOU$, and $MIOU$ evaluation indexes mentioned in Section 4.2 were adopted to evaluate the segmentation accuracy. The larger the value of the four indexes, the better the segmentation effect. The $IOU$ and $F_1$ test results for each category are shown in Table 3, and the $MIOU$ and $OA$ test results for all categories are shown in Table 4. The best results are marked in bold. It can be seen that our method achieves the largest $IOU$ and $F_1$ values in most categories and obtains the highest $MIOU$ and $OA$ in all categories. The segmentation effect of our method is the best in terms of the objective evaluation, which is consistent with the subjective evaluation. It should be noted that although our method has no significant improvement in objective results, it has a great improvement in subjective visual perceptions compared with other algorithms.

Table 5 shows the comparison results of parameters, FLOPs, and FPS between our method and other algorithms on the underwater dataset. It can be seen that our method has relatively

| Target category | Deeplab V3+ | DFANet | APCNet | Liu et al. | Our method |
|-----------------|-------------|--------|--------|------------|------------|
| Fish            | 63.4        | 75.2   | 60.2   | 77.2       | 66.1       | 76.3       | 64.4       | 76.9       | 70.5       | 78.0       |
| Diver           | 63.2        | 73.7   | 61.4   | 75.8       | 63.0       | 75.2       | 63.1       | 76.2       | 65.4       | 77.5       |
| Coral           | 61.1        | 74.1   | 66.2   | 75.4       | 63.5       | 76.4       | 63.8       | 76.5       | 65.6       | 76.3       |
| Rock            | 59.8        | 80.2   | 59.1   | 80.5       | 61.4       | 80.2       | 61.2       | 80.1       | 63.4       | 81.2       |
| Sculpture       | 62.4        | 74.5   | 62.1   | 76.1       | 64.2       | 78.4       | 63.0       | 77.8       | 69.5       | 77.4       |
| Octopus         | 64.6        | 72.3   | 61.8   | 75.1       | 63.6       | 75.2       | 66.3       | 74.9       | 67.1       | 75.7       |
| Turtle          | 67.0        | 75.9   | 60.3   | 77.7       | 63.7       | 77.1       | 65.8       | 77.8       | 66.7       | 78.6       |
| Seaweed         | 60.3        | 79.1   | 65.7   | 79.3       | 66.3       | 79.6       | 67.1       | 79.9       | 68.9       | 79.4       |
| Manta Ray       | 62.8        | 78.7   | 62.9   | 79.4       | 67.5       | 79.8       | 68.1       | 80.1       | 67.8       | 81.1       |
| Starfish        | 61.1        | 79.4   | 61.7   | 80.4       | 67.0       | 80.5       | 68.4       | 80.4       | 70.1       | 81.5       |
| Shell           | 60.2        | 75.1   | 60.9   | 75.7       | 64.1       | 76.3       | 62.6       | 76.5       | 65.9       | 77.3       |
| Sea urchin     | 60.7        | 79.5   | 64.5   | 80.6       | 64.9       | 80.7       | 63.4       | 80.7       | 67.8       | 82.0       |

https://doi.org/10.1371/journal.pone.0272666.t003
few parameters, small FLOPs, and less occupation of computing resources. Meanwhile, the operating speed is relatively fast, and the FPS reaches 125. Generally, our method has a good balance between segmentation accuracy and speed.

4.5 Ablation experiment

The contribution of each module in our method to algorithm performance was validated on the underwater image dataset. The ablation experiment results of adding modules DCHAC, ACSPP, CIA decoder, and UIE to the feature extraction network framework are presented in Table 6, where “Decoder” represents the decoder structure of DeeplabV3+. The experimental results show that the addition of each module has a certain effect, and the accuracy of image semantic segmentation has been improved. Among them, the addition of the UIE module has the greatest impact on the enhancement of underwater image semantic segmentation. And the objective evaluation index reaches the highest when the four modules are added, indicating that the four modules are indispensable, and the combined effect is the best for underwater image segmentation.

4.6 Failure examples

Fig 12 shows the failure examples of our proposed network. Since our segmentation network segments different objects by pixel region, some objects of different classes but similar shapes are difficult to distinguish. In addition, the change of scale leads to poor segmentation results for some small objects. As shown in Fig 12, the turtle of (a) is mistakenly detected as a rock due to the insignificant pixel difference, in (b) many little fish are not detected due to the light and shadow phenomenon caused by the refraction of the seawater, the humanoid sculpture in (c) was mistakenly detected as a diver, the octopuse in (d) is mistakenly identified as marine fish due to indistinguishable shapes.

5. Conclusion and future work

According to the characteristics of underwater images, this paper proposes an underwater image semantic segmentation method based on encoder-decoder architecture. Firstly, the image enhancement based on multi-spatial transformation is performed to enhance the original images. Then, the densely connected hybrid atrous convolution effectively expands the receptive field and slows down the speed of resolution reduction. Subsequently, the cascaded

| Method | Input size | Parameters(M) | FLOPs(G) | FPS |
|--------|------------|---------------|----------|-----|
| Deeplab V3+ | 513×513 | 213.1 | – | 0.4 |
| DFANet | 512×1024 | 7.8 | 1.7 | 160 |
| APCN | 512×1024 | – | – | – |
| CANet | 640×360 | 19.7 | 6.2 | 72 |
| Our method | 512×512 | 9.6 | 1.8 | 125 |

Table 5. Comparison of execution speed of different algorithms on the CUOID dataset.

https://doi.org/10.1371/journal.pone.0272666.t005
atrous convolutional spatial pyramid pooling module integrates boundary features of different scales to enrich target details. Finally, the context information aggregation decoder fuses the features of the shallow network and the deep network to extract rich contextual information, which greatly reduces information loss.

To verify the effectiveness of the proposed method, this paper established a comprehensive underwater image dataset, and our method was compared with four mainstream semantic segmentation methods through subjective and objective evaluations. The experimental results show that our method has obvious advantages in both subjective perception and objective data, and it outperforms the existing methods in terms of segmentation integrity, positioning accuracy, boundary clarity, and details. Meanwhile, ablation experiments verify the contribution of each component in our method to the final segmentation performance.

The method proposed in this paper improves the segmentation effect, but there are still misjudgments and omissions for objects with little discrimination. Our future work will focus on this problem to reduce the occurrence of misjudgments and omissions. Also, we will combine object detection and image enhancement to systematically analyze and improve underwater image processing.

**Acknowledgments**

The authors would like to thank all the reviewers who participated in the review, as well as MJEditor (www.mjeditor.com) for providing English editing services during the preparation of this manuscript.
Author Contributions
Conceptualization: Xiaohui He.
Data curation: Xiaohui He.
Formal analysis: Jinkang Wang, Xiaohui He.
Funding acquisition: Faming Shao.
Investigation: Faming Shao, Guanlin Lu, Ruizhe Hu.
Methodology: Jinkang Wang.
Project administration: Faming Shao.
Resources: Faming Shao, Qunyan Jiang.
Software: Guanlin Lu.
Supervision: Guanlin Lu, Ruizhe Hu.
Validation: Ruizhe Hu.
Visualization: Jinkang Wang, Qunyan Jiang.
Writing – original draft: Jinkang Wang.
Writing – review & editing: Xiaohui He, Qunyan Jiang.

References
1. Li C., Guo C., Ren W., Cong R., Hou J., Kwong S., et al. An underwater image enhancement benchmark dataset and beyond. IEEE Trans. Image Process. 2019, 29, 4376–4389. https://doi.org/10.1109/TIP.2019.2955241 PMID: 31796402
2. Li C., Saeed A., Fatih P. Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recog. 2020, 98, 107038.
3. Li Y., Chen R. UDA-Net: Densely attention network for underwater image enhancement. IET Image Process. 2021, 15, 774–785.
4. Wei X., Yu L., Tian S., Feng P., Ning X. Underwater target detection with an attention mechanism and improved scale. Multimedia Tools App. 2021, 80, 33747–33761.
5. Chen L.; Liu Z.; Tong L. Underwater object detection using Invert Multi-Class Adaboost with deep learning. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.
6. Fan Mingyuan, et al. "Rethinking BiSeNet for real-time semantic segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
7. Hua Xia, et al. "Cascaded panoptic segmentation method for high resolution remote sensing image." Applied Soft Computing 109 (2021): 107515.
8. Mahmmod Basheera M., et al. "Fast Computation of Hahn Polynomials for High Order Moments." IEEE Access 10 (2022): 48719–48732.
9. Abdulhussain Sadiq H., and Mahmmod Basheera M. "Fast and efficient recursive algorithm of Meixner polynomials." Journal of Real-Time Image Processing 18. 6 (2021): 2225–2237.
10. Guo Y.; Liu Y.; Georgiou T.; Lew M.S. A review of semantic segmentation using deep neural networks. Int. J. Multimedia Inform. Retr. 2018, 7, 87–93.
11. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man, Cyber. 1979, 9, 62–66.
12. Lafferty J.; Andrew M.; Fernando C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
13. Long J.; Evan S.; Trevor D. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2015).
14. Simonyan K.; Andrew Z. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014, arXiv:1409.1556.
15. Chen L.; Papandreu G.; Kokkinos I.; Murphy K.; Yuille A. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint, 2014, arXiv:1412.7062.
16. Chen L.; Papandreu G.; Kokkinos I.; Murphy K.; Yuille A. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Analy. Mach. Intel. 2017, 40, 834–848.
17. Chen L.; Papandreu G.; Schroff F.; Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint, 2017, arXiv:1706.05587.
18. Lin T.; Dollar P.; Girshick R.; He K.; Hariharan B.; Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2017).
19. Yu F.; Vladlen K.; Thomas F. Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2017).
20. He K.; Zhang X.; Ren S.; Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 37, 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824 PMID: 26353135
21. Yu F.; Vladlen K. Multi-scale context aggregation by dilated convolutions. arXiv preprint, 2015, arXiv:1511.07122.
22. Badrinarayanan V.; Alex K.; Roberto C. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intel. 2017, 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615 PMID: 28060704
23. Lin G.S.; Milan A.; Shenc H. Refinenet: multipath refinement networks for high-resolution semantic segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, (2017).
24. Zhang H.; Dana K.; Shi J.; Zhang Z. Context encoding for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, (2018).
25. Yang M.; Yu K.; Chi Z.; Li Z.; Yang K. Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2018).
26. Li H.; Xiong P.; Fan H.; Sun J. Difnet: Deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2019).
27. Fu J.; Liu J.; Tian H.; Li Y. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, (2019).
28. Liu C.; Chen L.; Schroff F.; Adam H.; Hua W. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2019).
29. Zoph B.; Vasudevan V.; Shlens J.; Le Q. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2018).
30. He J.; Deng Z.; Zhou L.; Wang Y.; Qiao Y. Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2019).
31. Zhang C.; Lin G.; Liu F.; Yao R.; Shen C. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2019).
32. Huynh Chuong, et al. "Progressive semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
33. Zheng Sixiao, et al. "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers." Computer Vision and Pattern Recognition IEEE, 2021.
34. Choi S., et al. "RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening." 2021.
35. Li S.; Xin G. Research of Underwater Image Segmentation Based on Improved Adaptive GACV Algorithm, 2018.
36. Yan Z.; Zhang J.; Yang Z.; Tang J. Kapur's entropy for underwater multilevel thresholding image segmentation based on whale optimization algorithm. IEEE Access 2020, 9, 41294–41319.
37. Chen W.; He C.; Ji C.; Zhang M.; Chen S. An improved K-means algorithm for underwater image background segmentation. Multimedia Tools App. 2021, 80, 21059–21063.
38. Ma W.; Xu F. Underwater image segmentation based on computer vision and research on recognition algorithm. Arabian J. Geosci. 2021, 14, 1–11.
39. Zhu Y.; Hao B.; Jiang B.; Rui N. Underwater image segmentation with co-saliency detection and local statistical active contour model. OCEANS 2017-Aberdeen. IEEE, 2017.
40. Liu F.; Fang M. Semantic segmentation of underwater images based on improved Deeplab. J. Marine Sci. Eng. 2020, 8, 188.
41. He K. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2016).
42. Huang G.; Liu Z.; Laurens V.; Weibecker K. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (2017).
43. Ancuti C. Color balance and fusion for underwater image enhancement. IEEE Trans. Image process. 2017, 27, 379–393. https://doi.org/10.1109/TIP.2017.2759252 PMID: 28981416
44. Chang Y.; Jung C.; Ko P.; Song H.; Hwang J. Automatic contrast-limited adaptive histogram equalization with dual gamma correction. Ieee Access 2018, 6, 11782–11792.
45. Ancuti C.; Ancuti C.; Haber T.; Bekaert P. Enhancing underwater images and videos by fusion. 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.
46. Tian H.; Xue M.; Ming Z.; Meng H. An improved multi-scale retinex fog and haze image enhancement method. 2016 International Conference on Information System and Artificial Intelligence (ISAI). IEEE, 2016.
47. He K.; Sun J.; Tang X. Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 35, 1397–1409.
48. Wang P.; Chen P.; Yuan Y.; Liu D. Understanding convolution for semantic segmentation. 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.
49. Rivero-Castillo Daniel, Pijeira Héctor, and Pedro Assunção. "Edge detection based on Krawtchok polynomials." Journal of computational and applied mathematics 284 (2015): 244–250.
50. Abdulhussain Sadiq H., et al. "Image edge detection operators based on orthogonal polynomials." International Journal of Image and Data Fusion 8. 3 (2017): 293–308.
51. Lin T.; Maire M.; Belongie S.; Hays J.; Zitnick C. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, Cham, 2014.
52. Everingham M.; Eslami S.; Van Gool L.; Williams C. The pascal visual object classes (voc) challenge. Int. J Comp. Vis. 2010, 88, 303–338.
53. Liu R.; Fan X.; Zhu M.; Hou M.; Luo Z. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. IEEE Trans. Circuits Syst. Video Tech. 2020, 30, 4861–4875.
54. Northeast Fisheries Science Center. Habitat mapping camera(Habcam), 2012.
55. Ayob A. F., et al. "Analysis of pruned neural networks (MobileNetV2-YOLO v2) for underwater object detection." Proceedings of the 11th National Technical Seminar on Unmanned System Technology 2019. Springer, Singapore, 2021.
56. Drews-Paulo Jr, et al. "Underwater image segmentation in the wild using deep learning." Journal of the Brazilian Computer Society 27. 1 (2021): 1–14.