Robotica (2023), **41**, pp. 3541–3564
doi:10.1017/S026357472300098X

RESEARCH ARTICLE

Global enhancement network underwater archaeology scene parsing method

Junyan Pan[1], Jishen Jia[1,2] and Lei Cai[3]

[1]School of Mathematical Sciences, Henan Institute of Science and Technology, Xinxiang, China, [2]Henan Digital Agriculture Engineering Technology Research Center, Xinxiang, China, and [3]School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang, China

Corresponding author: Lei Cai; Email: cailei2014@126.com

Received: 30 October 2022; **Revised:** 30 April 2023; **Accepted:** 12 June 2023; **First published online:** 8 September 2023

Keywords: underwater archaeology; target semi-buried; adaptive dilated convolution; feature fuzziness; enhancement network

Abstract

Underwater archaeology is of great significance for historical and cultural transmission and preservation of underwater heritage, but it is also a challenging task. Underwater heritage is located in an environment with high sediment content, objects are mostly buried, and the water is turbid, resulting in some of the features of objects missing or blurred, making it difficult to accurately identify and understand the semantics of various objects in the scene. To tackle these issues, this paper proposes a global enhancement network (GENet) underwater scene parsing method. We introduce adaptive dilated convolution by adding an extra regression layer, which can automatically deduce adaptive dilated coefficients according to the different scene objects. In addition, considering the easy confusion in the process of fuzzy feature classification, an enhancement classification network is proposed to increase the difference between various types of probabilities by reducing the loss function. We verified the validity of the proposed model by conducting numerous experiments on the Underwater Shipwreck Scenes (USS) dataset. We achieve state-of-the-art performance compared to the current state-of-the-art algorithm under three different conditions: conventional, relic semi-buried, and turbidified water quality. The experimental results show that the proposed algorithm performs best in different situations. To verify the generalizability of the proposed algorithm, we conducted comparative experiments on the current publicly available Cityscapes, ADE20K, and the underwater dataset SUIM. The experimental results show that this paper achieves good performance on the public dataset, indicating that the proposed algorithm is generalizable.

1. Introduction

The documentation of underwater cultural heritage is the basis for sustainable marine development, and underwater archaeology is great significance for historical and cultural transmission and preservation of underwater heritage. The purpose of scene resolution is to assign a class label to each pixel in the image, and effective scene resolution can provide important meaningful values for underwater archaeology.

In reality, in classical FCN [1] networks, high-precision parsing of global and local information usually cannot be achieved simultaneously. To improve the information extraction of global context, Jie Jiang et al. [2] proposed a novel global-guided selective context network (GSCNet) to select contextual information adaptively. Multi-scale feature fusion and enhancement network (MFFENet) [3], semantic consistency module [4], and attention residual block-embedded adversarial networks (AREANs) [5] can fuse global semantic information at multiple scales. For information extraction of local information features, Shiyu Liu et al. [6] highlighted the strong correlation between depth and semantic information by introducing a built-in deep semantic coupling coding module that adaptively fuses RGB and depth features. Junjie Jiang et al. [7] proposed a graphical focus network that captures the local detectors of objects and relational dependencies. Zhitong Xiong [8] et al. adopted a novel variational context

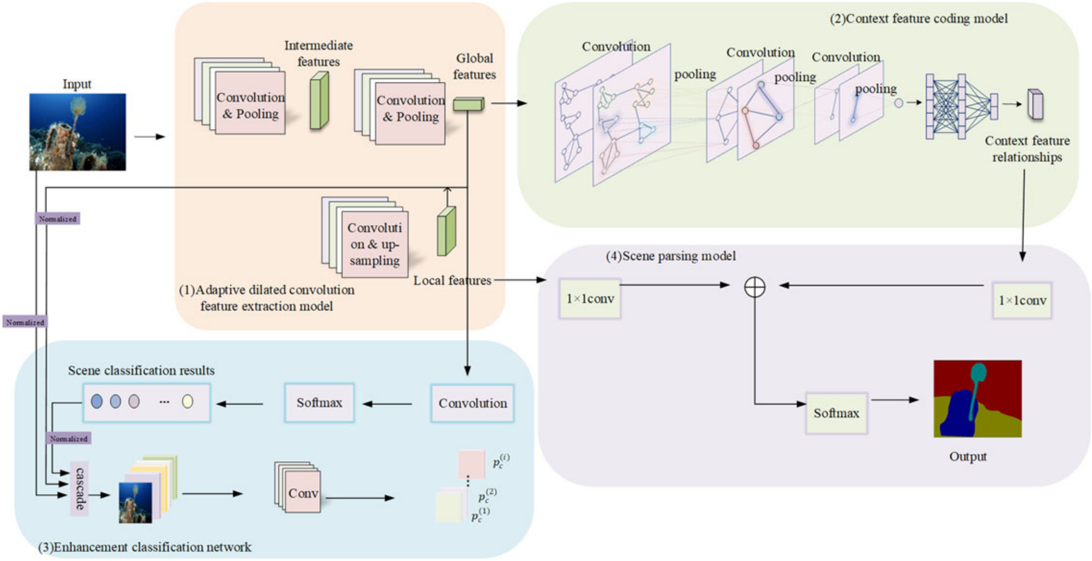


Figure 1. *Overview of the proposed method.*

deformable module that learns adaptive environmental changes in a structured way to improve the recognition of local information. However, the existing methods are not able to decompose spatial details and contextual information learning better, and the accuracy of resolution for fuzzy features is not high.

To address the above issues, this paper proposes a global enhancement network underwater scene parsing method, as shown in Fig. 1. Our main contributions can be summarized as follows:

1. The receptive field in standard convolution is fixed and can only capture a uniform feature scale. However, underwater artifacts have different shapes and sizes. To extract features of multiple scales of target classes in underwater archaeological scenes, we propose adaptive dilated convolution. It can learn the dilated coefficients adaptively according to the size of the objects in the scene and can capture global features more flexibly and efficiently so that features with various sizes in the scene can be handled effectively.
2. The image features of the data collected under turbid water conditions are blurred, and different categories are easily confused in the process of classification. Traditional classification methods use a single classifier, which has little discrimination between features with high similarity. To improve the category classification accuracy, this paper introduces multiple classifiers and uses multiple convolutional layers for enhanced classification. A difference-based regularization method is used for feature categories with high similarity to enhance the probabilistic scoring differences between categories and improve classification accuracy.
3. To verify the effectiveness and advance of the proposed method in solving Underwater archaeology problems, we self-made the Underwater Shipwreck Scenes (USS) dataset and conducted experimental comparison between the current advanced method and the proposed algorithm on the USS dataset. Also, to verify the generalization of the proposed algorithm, we compare the proposed algorithm with the current state-of-the-art algorithm on general datasets (ADE20K and Cityviews) and underwater datasets (SUIM). The experimental results show that the proposed algorithm achieves good performance on different datasets.

2. Related work

In scene parsing, significant results have been achieved recently for the problem of missing some features due to the occlusion of the target. Some approaches [9, 10] build overall contextual relationships

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

based on fully convolutional neural networks to capture more detailed features and thus improve the parsing performance. High-level semantic features in deep feature aggregation networks (DFANet) [11] can abstract effective features by successively combining low-level detailed features, capable of simultaneous large sensory fields and detailed spatial features. DecoupleSegNets [12] improves the semantic segmentation performance by modeling the target body and edge explicitly. Semantic-aware occlusion robust network [13] used the intrinsic relationship between the recognition target and the occluded part to infer the missing features. Semantic guidance and estimation network (SeGuE-Net) [14] can reconstruct and repair the missing data. L. Cai [15] proposed an enhanced dilated convolution framework for underwater blurred target recognition. X. Qiao et al. [16] adopted a deep neural network to solve the problem of phantom missing in the attributes of independent objects contextual information problem. Context-based tandem network (CTNet) [17] can effectively combine global and local information to improve the performance of semantic segmentation. Semantic structure aware [18] proposes a semi-supervised semantic segmentation algorithm based on semantic structure awareness. By exploiting the relationship between different semantic structures in the training data, the weakly supervised information can be transformed into strongly supervised information, thus improving the pixel-level dense prediction accuracy without increasing the cost. Dilated convolution with learnable spacings (DCLS) [19] can increase the size of the perceptual field without increasing the parameters. The interpolation technique is used to flexibly determine the spacing between non-zero elements, or equivalent positions, by back-propagation learning. Adaptive fractional dilated convolution network (AFDC) [20] with aspect ratio embedding, component preservation, and parameter-free features. The method adaptively constructs fractional-order dilated kernels based on the aspect ratio of the image and uses the two closest integer-order dilated kernels to interpolate to solve the misalignment problem of fractional-order sampling. Adaptive dilated convolution (ADC) [21] can generate and fuse multi-scale features of the same spatial size by setting different dilated rates for different channels. It enables the ADC to adaptively adjust the fusion scale to better fit the various sizes of the target class. However, the receptive field size of the above model is fixed, and only a uniform feature proportion can be captured for objects with inconsistent size.

Deep water environments are poorly lit and sediment is present, making it difficult to accurately identify features. To improve the semantic performance of degraded images, X. Niu et al. [22] proposed an effective image recovery framework based on generative adversarial networks. The underwater distorted target recognition network (UDTRNet) [23] and the method of binary cross-entropy loss to extract abstract features of objects [24] improved the detection accuracy of underwater targets. Object-guided dual-adversarial contrast learning [25] and multi-scale fusion algorithm [26] can effectively enhance seriously distorted underwater images. Zhi Wang et al. [27] proposed an adaptive global feature enhancement network (AGFE-Net) that used multi-scale convolution with global receptive fields and attention mechanisms to obtain multi-scale semantic features and enhance the correlation between features. Asymmetric non-local neural networks for semantic segmentation (ANNNet) [28] designed an asymmetric Non-Local approach to computing point-to-point similarity relationships efficiently and aggregating global information and context. W. Zhou et al. [29] proposed a common extraction and gate fusion network (CEGFNet) for capturing high-level semantic features and low-level spatial details for scene parsing. SegFormer [30] is capable of outputting multi-scale features and aggregating information from different network layers. Y. Sun et al. [31] proposed a model fine-tuning method based on singular value decomposition, which can perform fast model fine-tuning by using a very small number of parameters, thus achieving the segmentation task with few samples. Z. Li et al. [32]proposed a knowledge-guided approach for few-sample image recognition. A knowledge-guided model based on an attention mechanism is used to achieve the classification of few-sample images in the target domain by fusing and utilizing knowledge from multiple source domains. Semantic Segmentation of Underwater (SUIM Net) proposed by M. J. Islam et al. [33], improves the performance of semantic segmentation using a fully convolutional encoder-decoder model. Several studies [34–36] combined conditional random fields with deep CNNs, with significant improvements in segmentation accuracy and generalization performance. These scene resolution methods can effectively enhance image features, but ignore the similarity of fuzzy features in the classification process is prone to misclassification.

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

In this paper, we propose adaptive dilated convolutional networks with flexible size and shape receptive fields so as to tackle objects of different sizes in the scene. We also propose an enhancement classification network that can effectively improve the classification accuracy of fuzzy features. In addition, the proposed method in this paper can effectively capture both global and local contextual information.

3. Our approach

The proposed method contains four parts, as shown in Fig. 1. The first part is adaptive dilated convolution feature extraction model P_1. The global and local features of the scene are extracted by the method of adaptive dilated convolutional network. The second part is the contextual feature encoding model P_1. This part is based on the overall scene and regional features to learn their contextual relationships. The third part enhancement classification model. For confusable objects, an enhancement classifier is used to discriminate the correct object class. The fourth part is the scene parsing model. Based on global and local features and contextual relationship features, classification is performed to get the final scene parsing results.

3.1. Adaptive dilated convolution feature extraction model

Let the input features be $X^{(i)} \in \mathbb{R}^{H\times W\times D}$, where $i = 1, \ldots, N$, N is the total number of scene elements in the dataset; H and W are the height and width of the image, respectively; D is the number of channels of the image. Firstly, we need to extract the features from the input elements $X^{(i)}$. The feature extraction model is as follows:

$$\begin{gathered} \left\{\alpha^{(i)}, \beta_1^{(i)}, \beta_2^{(i)}, \ldots, \beta_J^{(i)}\right\} = P_1\left(X^{(i)}; \xi_1\right) \\ \alpha^{(i)} \in \mathbb{R}^{D_\alpha} \\ \beta_j^{(i)} \in \mathbb{R}^{D_r, j} = 1, \ldots, J \end{gathered} \tag{1}$$

where $\alpha^{(i)}$ is the feature vector of D_α dimension, which encodes the semantic information of the overall scene. $\left\{\beta_j^{(i)}\right\}_{j=1,\ldots,J}$ is the feature vector of D_β dimension, which encodes the local semantic information. J is the number of feature regions in the scene, and ξ_1 is the parameter of feature encoding model P_1.

In the process of underwater archaeology, underwater heritage exists buried by mud and sand, and some features are missing, which brings great challenges to the extraction of features. Extracting the overall scene features $\alpha^{(i)}$ and local features $\left\{\beta_j^{(i)}\right\}_{j=1,\ldots,J}$ is a prerequisite for generating accurate scene parsing. Convolutional neural network (CNN) can extract high-level semantic features better. However, in the standard convolution, the field of receptive is fixed, and features with different sizes of objects cannot be extracted simultaneously. In contrast, dilated convolution adjusts the size and shape of the convolution block by using the learned dilated coefficients, and has a flexible size and shape of the receptive field. At the same time, multi-scale feature information is extracted without loss of spatial resolution. To extract both overall and local features, we propose an adaptive dilated convolutional network. In the paper, the proposed method removes the two pooling layers in the convolutional neural network to obtain more detailed local features, to obtain higher-resolution feature maps. This paper adaptively adjusts different dilated coefficients according to task requirements. For small target classes requiring detailed feature learning, a smaller dilated coefficient (less than 1) is used to learn more detailed local features $\left\{\beta_j^{(i)}\right\}_{j=1,\ldots,J}$. For large target classes that require global features, use a larger dilated coefficient (greater than 1) to learn the broader overall feature $\alpha^{(i)}$.

Figure 2 provides the overview of the proposed adaptive dilated convolution. Assume that the input and output features of the dilated convolution are $X^{(i)} \in \mathbb{R}^{H\times W\times D}$ and $F^{(i)} \in \mathbb{R}^{H\times W\times D}$, respectively. The dilated coefficients are learned from an additional regression layer that takes the feature map X as input and outputs an dilated coefficient map $\mathrm{R} \in \mathbb{R}^{H\times W\times 8}$. The size and shape of receptive field are adjusted by the vector coefficients in R. Since the position of the dilation coefficient vectors is adaptive, the dilated

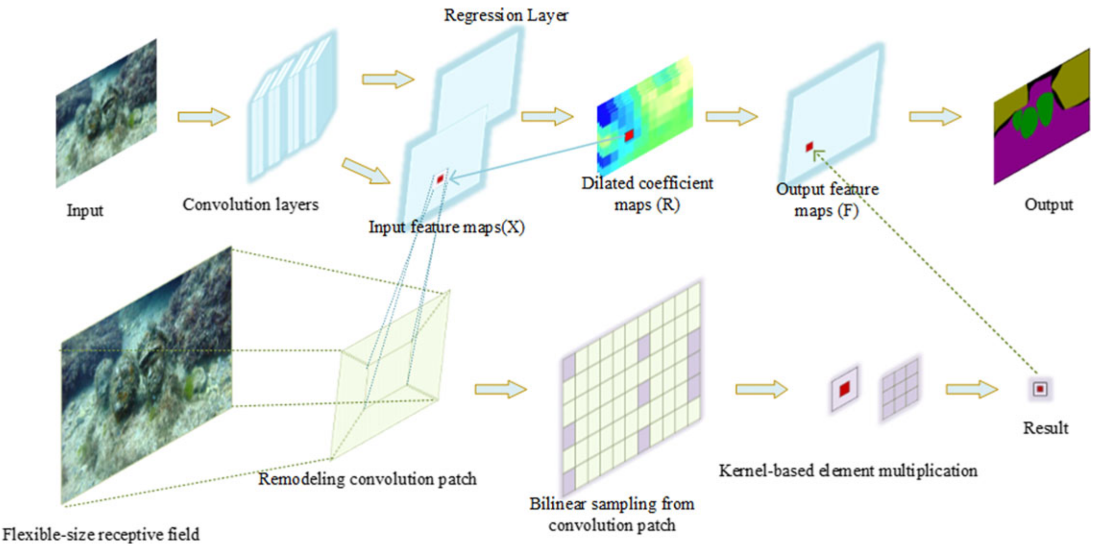

Regression Layer
Input
Convolution layers
Input feature maps(X)
Dilated coefficient maps (R)
Output feature maps (F)
Output
Flexible-size receptive field
Remodeling convolution patch
Bilinear sampling from convolution patch
Kernel-based element multiplication
Result

Figure 2. *Overview of adaptive dilated convolution.*

coefficient maps R has the same size as the output feature maps F. Thus, all the convolution patches have their respective dilated coefficient vectors to obtain the receptive field of the target size and shape.

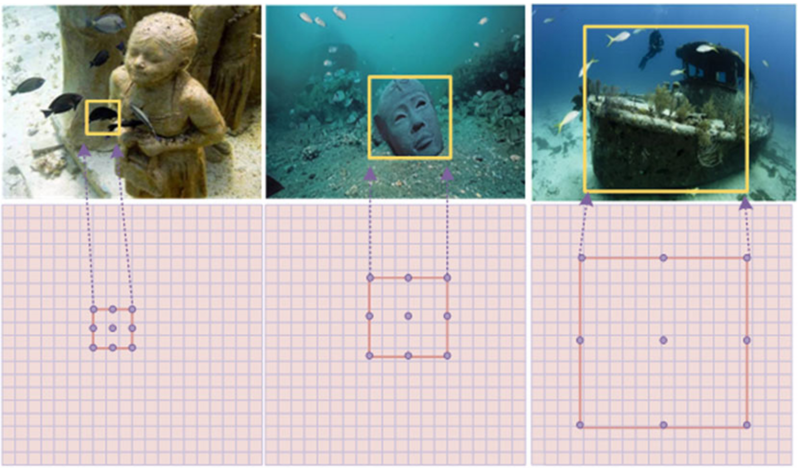
We use formula (2) to initialize the dilated regression layer:

$$\begin{cases} \eta_0(a) = \varepsilon \\ \mu_0 = 1 \\ \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right), \quad \sigma \ll 1 \end{cases} \tag{2}$$

where η_0 and μ_0 denote the initialized deviations of the convolution kernel and the regression layer, and a denotes the position in the convolution kernel. The convolution kernel is set to a value close to 0 and the bias is 1, so the generated dilated coefficients are almost approach to 1. The training starts with the standard convolution, and the adaptive dilated coefficients are progressively learned during the training process to obtain the appropriate values.

Reconstructing convolutional patches extracts a set of patches from the feature map, each of which has a certain shape and size. The number, size, and shape of the convolution patches are related to the size of the input feature map and the receptive field size of the network, which are adaptively reshaped based on the dilated coefficient R. These patches are then reconstructed into a new feature map by interpolating between the patches to fill in the missing pixels. (The missing pixels in the patches refer to the areas in the reconstructed feature map that may have gaps or pixel values that were not directly observed or sampled. These missing pixels occur because the patches are extracted from the original feature map, and during the reconstruction process, interpolation is used to fill in the gaps between the patches). This approach allows the network to consider larger scenes when performing convolutional operations and can increase the perceptual field of the feature map. In traditional convolutional neural networks, reducing the size and resolution of the feature maps by using operations such as pooling layers and step convolution can lead to loss of information and reduced resolution. To overcome this problem, a bilinear sampling approach is used. The feature vectors are then sampled from the convolution patch in a bilinear interpolation to perform element-by-element multiplication with the kernel. We apply the adaptive dilated convolution to the last layer of the CNN. For each location (the red dot as an example), the associated convolutional patch is reshaped using the off-dilated coefficient vector learned from the regression layer to obtain a perceptual field with flexible size and shape. When the dilated coefficient is equal to 1, the adaptive convolution is the standard convolution; when the dilated coefficient is less

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press



(1)Coefficients < 1 (2) Coefficients = 1 (3) Coefficients > 1

Figure 3. *Illustration of receptive fields with flexible sizes in adaptive dilated convolutions.*

than 1, the convolution patch shrinks and the perceptual field shrinks; when the dilated coefficient is greater than 1, the convolution patch expands so that the perceptual field is enlarged. Figure 3 illustrates receptive fields with flexible sizes learned from dilated convolution. We perform experiments using different sizes of convolution kernels, and the experimental results show that 1×1, 3×3, 5×5, and 7×7 convolution kernels perform the best convolution, and other sizes lose some of their ability to convolve feature information at different scales. Here, the convolution kernels are operated for different perceptual fields rather than changing the size of the kernel by applying an expansion factor. These convolution kernels are of fixed size and are used to capture feature information at different scales under different receptive fields. For each receptive field, they capture local details, medium-scale structures, and larger-scale contextual information, respectively.

Let R^t be the 8-dimensional vector of the adaptive dilated coefficient at position t, $R^t = \left[r^t_{x1}, r^t_{x2}, r^t_{x3}, r^t_{x4}, r^t_{y1}, r^t_{y2}, r^t_{y3}, r^t_{y4}\right]^T$. Suppose that for F^t, the associated convolution patch in X is W^t, and its center (s^t, v^t) is any convex quadrilateral, the convolution kernel is $K \in \mathbb{R}^{D\times(2k+1)\times(2k+1)}$, where k is a constant. The shape of W^t is determined by R^t by changing the position of its four corners, each of which is controlled by two components of R^t, and the values of all components in R^t are determined by the position of the corners. Feature vectors $(2k+1)\times(2k+1)$ were selected from the convolution patch W^t for element multiplication. The coordinates of these feature vectors $\left(x_{ij}, y_{ij}\right)$ were determined by the new positions of corner points, and it can be expressed as:

$$\begin{cases} x_{ij} = s^t + \varphi ij\dfrac{r_{x1}+r_{x2}-r_{x3}-r_{x4}}{4k} + \varphi i\dfrac{r_{x1}+r_{x2}+r_{x3}+r_{x4}}{4} \\ \qquad + \varphi j\dfrac{r_{x1}-r_{x2}+r_{x3}-r_{x4}}{4} + \varphi k\dfrac{r_{x1}-r_{x2}-r_{x3}+r_{x4}}{4} \\ y_{ij} = v^t + \varphi ij\dfrac{r_{y1}+r_{y2}-r_{y3}-r_{y4}}{4k} + \varphi i\dfrac{r_{y1}+r_{y2}+r_{y3}+r_{y4}}{4} \\ \qquad + \varphi j\dfrac{r_{y1}-r_{y2}+r_{y3}-r_{y4}}{4} + \varphi k\dfrac{r_{y1}-r_{y2}-r_{y3}+r_{y4}}{4} \end{cases} \tag{3}$$

where s^t, v^t, φ are integers, $i, j \in [-k, k]$.

Since the dilated coefficient is a real value, the eigenvector can be obtained by bilinear interpolation. Suppose the convolution patch after bilinear interpolation is O_t, and it can be expressed as:

$$O^t_{ij} = \sum_{n,m} W^t_{nm} \max\left(0, 1-\left|x''_{ij}-m\right|\right) \max\left(0, 1-\left|y''_{ij}-n\right|\right) \tag{4}$$

where $W^t_{nm} = W^t(n, m)$, n, m $\in [-\varphi$k, φk]. So the forward propagation of convolution is:

$$F^t = \sum_{i,j} K_{ij} O^t_{ij} + \mu \tag{5}$$

where $K_{ij} = K(i, j)$ denotes the multiplication process of all output channel elements and μ is the bias.

Accordingly, in the case of backpropagation, the gradient change is:

$$\begin{aligned} g\left(O^t_{ij}\right) &= \left(K_{ij}\right)^T g\left(F^t\right) \\ g\left(K_{ij}\right) &= g\left(F^t\right)\left(O^t_{ij}\right)^T \\ g(\mu) &= g\left(F^t\right) \end{aligned} \tag{6}$$

We obtain $\frac{\partial O^t_{ij}}{\partial W^t_{ij}}$ for the bilinear interpolation equation and then obtain $\frac{\partial o^t_{ij}}{\partial x_{ij}}$ and $\frac{\partial O^t_{ij}}{\partial y_{ij}}$ for the partial derivatives of the corresponding coordinates. Since the coordinates x_{ij} and y_{ij} depend on the vector of perspective coefficients R^t, we can use the following partial derivatives of the perspective coefficients to obtain the gradient of R^t.

$$\begin{gathered} \begin{bmatrix} \partial x_{ij}/\partial r^t_{x1} \\ \partial x_{ij}/\partial r^t_{x2} \\ \partial x_{ij}/\partial r^t_{x3} \\ \partial x_{ij}/\partial r^t_{x4} \end{bmatrix} = \begin{bmatrix} ij\varphi + ik\varphi + jk\varphi + k^2\varphi \\ ij\varphi + ik\varphi - jk\varphi - k^2\varphi \\ -ij\varphi + ik\varphi + jk\varphi - k^2\varphi \\ -ij\varphi + ik\varphi - jk\varphi + k^2\varphi \end{bmatrix} /4k, \\ \begin{bmatrix} \partial y_{ij}/\partial r^t_{y1} \\ \partial y_{ij}/\partial r^t_{y2} \\ \partial y_{ij}/\partial r^t_{y3} \\ \partial y_{ij}/\partial r^t_{y4} \end{bmatrix} = \begin{bmatrix} ij\varphi + ik\varphi + jk\varphi + k^2\varphi \\ -ij\varphi - ik\varphi + jk\varphi + k^2\varphi \\ -ij\varphi + ik\varphi + jk\varphi - k^2\varphi \\ ij\varphi - ik\varphi + jk\varphi - k^2\varphi \end{bmatrix} /4k \end{gathered} \tag{7}$$

Finally, using these partial derivatives, the gradients of the perspective coefficient mapping R and the input feature mapping X are obtained by the chain rule.

$$\begin{aligned} g\left(r^t_{x\_}\right) &= \sum_{i,j} \left(\frac{\partial x_{ij}}{\partial r^t_{x\_}} \frac{\partial V^t_{ij}}{\partial x_{ij}} \right)^T g\left(O^t_{ij}\right), \\ g\left(r^t_{y\_}\right) &= \sum_{i,j} \left(\frac{\partial y_{ij}}{\partial r^t_{y\_}} \frac{\partial V^t_{ij}}{\partial y_{ij}} \right)^T g\left(O^t_{ij}\right), \\ g\left(W^t_{nm}\right) &= \sum_{i,j} \frac{\partial O^t_{ij}}{\partial W^t_{nm}} g\left(O^t_{ij}\right), \end{aligned} \tag{8}$$

where $r^t_{x\_}$ denotes any component of $\left\{r^t_{x1}, r^t_{x2}, r^t_{x3}, r^t_{x4}\right\}$ and $r^t_{y\_}$ denotes any component of $\left\{r^1_{y\_}, r^t_{y_2}, r^t_{y_3}, y^t_{x_4}\right\}$.

3.2. Context feature coding model

There are interconnections between scene features, and each scene feature is related to the overall scene, and these associations help feature classification and generate more accurate scene parsing results. By contextual feature encoding model P_2, the contextual association features between scene as a whole and localities are learned. As shown in equation (6):

$$\begin{gathered} \left\{\gamma^{(i)}_1, \gamma^{(i)}_2, \ldots, \gamma^{(i)}_J\right\} = P_2\left(\alpha^{(i)}, \beta^{(i)}_1, \beta^{(i)}_2, \ldots, \beta^{(i)}_J; \xi_2\right) \\ \gamma^{(i)}_i \in \mathbb{R}^{D_\gamma}, j = 1, \ldots, J \end{gathered} \tag{9}$$

where ξ_2 is the parameter of the contextual feature encoding model. $\left\{\gamma^{(i)}_j\right\}_{j=1,\ldots,J}$ is the D_γ-dimensional vector of contextual features. Each contextual feature $\gamma^{(i)}_i$ encodes the semantic relationship between the jth feature and the whole scene as well as the relationship between the jth feature and other features.

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

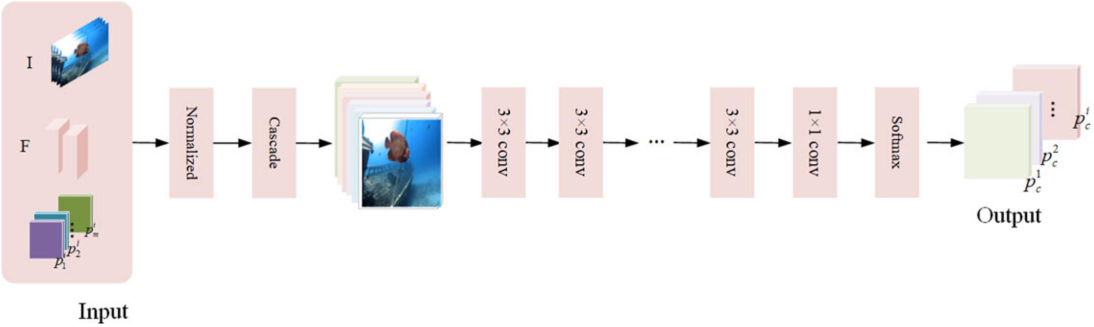


Figure 4. *The general diagram of enhancement classification network.*

The structure of graph is irregular and has infinite dimensionality, which can describe the semantic contextual relationship between feature points well, so graph convolutional network is used to learn the contextual relationship of scene features. The matrix E is used to represent the nodes of the scene feature graph, A is the adjacency matrix to represent the feature semantic contextual relationship, and the matrices E and A are the inputs of the graph convolutional network. The propagation between each layer in the graph convolutional neural network can be expressed as:

$$F^{k+1} = f\left(\tilde{D}^{-\frac{1}{2}}\tilde{E}\tilde{D}^{-\frac{1}{2}}F^k\rho^k\right) \tag{10}$$

where $\tilde{A} = A + I$, where I is the unit matrix, $\tilde{D}$ is the degree matrix of $\tilde{E}$, F^{k+1} is the feature of $k+1$ layers, the feature representation of the input layer X is E, $f(\cdot)$ is the nonlinear activation function, and ρ^k is the parameter of learning. The final output of the graph convolutional network is the updated features of node $X^{(i)}$ in the graph, which can be aggregated into a scene feature vector for feature semantic relationship inference.

3.3. Enhancement classification network

Let the input of scene classification model P_3 be scene features $\alpha^{(i)}$, and the output scene classification probability prediction result can be expressed as:

$$\begin{gathered}\left\{p_1^{(i)}, p_2^{(i)}, \ldots, p_M^{(i)}\right\} = P_3\left(\alpha^{(i)}; \xi_3\right) \\ p_m^{(i)} \in [0, 1], \quad m = 1, \ldots, M\end{gathered} \tag{11}$$

where ξ_3 is the parameter of the scene classification model, M is the total number of scene classes in the dataset, and $p_m^{(i)}$ is the initial probability that the image $X^{(i)}$ predicted by the scene classification model belongs to the mth class of scenes. The network includes a convolution layer with a filter size of 1 × 1, and a softmax regression layer, which can generate scene classification results $\left\{p_m^{(i)}\right\}$ $m = 1, \ldots, H \times W$ based on the input scene global feature $\alpha^{(i)}$.

The underwater heritage is located in an environment with turbid water, and the acquired object features are blurred, which can easily confuse the objects in the process of classification. The reinforced classification network is mainly used to discriminate the categories with similar probability values, from which the correct object category is judged. The input of the enhanced classification network is the initial probability value of $p_m^{(i)}$, the image features F and the original image X cascaded together. where the image features F are used as a reference during the enhancement classification to improve the enhancement classification accuracy. Since the initial probability $p_m^{(i)}$, the image features F and the original image X may have a wide range of values, these features are normalized separately. The general diagram of enhancement classification network is shown in Fig. 4.

Based on this input, a series of convolutional layers are then used for enhanced classification. First, a 3×3 convolution is used, which extracts information about the contextual relationships between multiple pixels. We introduce a multi-classifier here and then use a 1×1 convolution to predict the enhanced classification $\alpha_c^{(i)} \in \mathbb{R}^{H\times W\times D}$. Each of these elements $\alpha_c^{(i)}$ represents the probability value that the ith pixel in the image belongs to the cth object class. This probability value is normalized by the multinomial logistic regression (i.e., softmax) function as follows:

$$p_c^{(i)} = \frac{\exp\left(\alpha_c^{(i)}\right)}{\sum_{j=1}^{C} \exp\left(\alpha_j^{(i)}\right)} \tag{12}$$

where $p_c^{(i)}$ is the normalized probability. The set of all class probabilities $\left\{p_c^{(i)}\right\}_{C=1,\ldots,C}$ is the probability distribution $P(Y \mid X, c)$ sought by the scene semantic parsing task.

In the model training phase, after generating scene classification results, the scene classification loss is calculated. The loss of the strengthened classifier is the multiclassification cross-entropy loss function:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \times \log p_c^{(i)} \tag{13}$$

where $y_{i,c} \in \{0, 1\}$ is the scene classification label. $y_{(i,c)} = 1$ means that the image $X^{(i)}$ belongs to the cth category scene, and $y_{(i,c)} = 0$ means vice versa. Since the similarity of probability scoring of multiple categories can easily lead to misclassification, regularization is used to avoid this situation. The difference-based regularization method makes each category probability scoring as large as possible. Here, second-order moments are used as the difference-based regularization:

$$Q = 1 - \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \left(p_c^{(i)}\right)^2 \tag{14}$$

where $Q \in [0, 1 - 1/C]$. The larger the difference between the probability scoring $\left\{p_c^{(i)}\right\}_{C=1,\ldots,C}$, the smaller the value of Q. Thus, the loss function can be reduced to increase the difference between the probabilities.

3.4. Scene parsing model

To generate the scene semantic parsing results, this chapter adopts the scene parsing model P_4 to predict the category of each region based on the local features and contextual relationship features, so as to obtain the final scene parsing results. Since there are two input features, local features and contextual features, the scene parsing model P_4 first generates two sets of classification probabilities based on these two features and then integrates the two sets of probabilities. To obtain more detailed predictions, the scene resolution model P_4 predicts a set of classification probabilities based on the pixel-level local features $\left\{\beta_{j'}^{(i)}\right\}_{t'=1,\ldots,J'}$, where J' is the number of pixels in the image $X^{(i)}$. And another set of classification probabilities are predicted based on the hyperpixel-level contextual relationship features $\left\{\gamma_j^{(i)}\right\}_{j=1,\ldots,J}$, where J is the number of hyperpixels. After converting the probability results predicted based on the hyperpixel-level contextual relationship features to pixel-level results, they are then integrated with the results predicted based on the pixel-level features. The integration is done by taking the higher of the two probability scores for each pixel for each category as the final classification probability.

The scenario parsing model P_4 can be modeled as follows:

$$\begin{gathered}\left\{q_{j',1'}^{(i)}, q_{j',2}^{(i)}, \ldots, q_{j',L^{(i)}}^{(i)} = P_4\left(\beta_{j'}^{(i)}, \gamma_j^{(i)}; \xi_4\right)\right. \\ q_{j',k}^{(i)} \in [0, 1],\ \ k = 1, \ldots, K\end{gathered} \tag{15}$$

where ξ_4 is the parameter of the scene resolution model, K is the total number of object classes in the dataset, and $q_{j',k}$ is the probability that the j'th pixel in the image $X^{(i)}$ predicted by the scene resolution model belongs to the kth class of objects. In the model training phase, the scene resolution loss is used to constrain the learning of the model. The scene parsing loss uses a pixel-level cross-entropy loss function:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{J'}\sum_{j'=1}^{J'}\sum_{k=1}^{K} y_{i,j',k} \times \log q_{t',k}^{(i)} \tag{16}$$

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

where $y_{i,j',k}$ is the label of whether the j'th pixel belongs to the kth class of objects. Ultimately, a joint optimization approach can be used to combine the scene classification loss function and the scene resolution loss function, while optimizing all the models proposed in this chapter:

$$\begin{aligned} L &= L_1 + \lambda L_2 \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c=1}^{C} y_{i,c} \times \log p_c^{(i)}\right. \\ &\quad \left. +\lambda \frac{1}{J'}\sum_{t'=1}^{J'}\sum_{k=1}^{K} y_{i,j',k} \times \log q_{j',k}^{(i)}\right] \end{aligned} \tag{17}$$

where λ is a scaling factor that controls the ratio of scene classification loss to scene parsing loss.

4. Experiment and analysis

4.1. Training dataset and evaluation index

In order to evaluate the scene resolution performance of the proposed methods in this paper, the test results are compared with the current state-of-the-art methods. All methods are tested on the Underwater Shipwreck Scenes (USS) dataset, which contains 1163 images of eight categories: wreck, statue, porcelain, sediment, reef, water, plant, and fish. The data of this dataset are a homemade dataset generated by collecting underwater archaeology images online and then by manual annotation. Pixel accuracy (PAcc) and mean intersection-over-union accuracy (mIoU) are used as the evaluation criteria of the proposed algorithm in this paper. Pixel accuracy counts the percentage of correctly classified pixels in the whole dataset. The mean intersection-over-union (mIoU) is calculated separately for each object class, and then, the intersection-over-union values of all object classes are averaged as the final result.

4.2. Experiment details

In this experiment, training and testing are performed on a small server with GTX2080 GPU and 64G RAM. In order to demonstrate the objectivity of the proposed method in the comparison experiments, the Pytorch deep learning tool is used for implementation, and the ResNet convolutional neural network is used for deep network construction. The experimental training parameters are: initial learning rate = 0.0001, epoch = 300, momentum = 0.9, weight decay = 0.0001, and random gradient descent method is used for training. The learning rate is dynamically adjusted using the "poly" strategy.

4.3. Ablation experiment

We validate the effectiveness of adaptive dilated convolution (ADC) and enhancement classification network (ECN) in the proposed algorithm in this paper by ablation experiments. The experiments are conducted on the USS dataset. We use ResNet-50 as the baseline for global enhancement network (GENet) ablation experiments. The experimental results are shown in Table I and include the recognition accuracy of each category in addition to PAcc and mIoU. The visual results are shown in Fig. 5. From Table I, it can be seen that PAcc and mIoU in the baseline network, reached 79.6% and 63.8%. After adding ADC to the baseline, PAcc and mIoU are improved, which are 82.3% and 65.1%, respectively. Some categories such as "statue," "fish," and "plant" far exceeded the baseline. This indicates that ADC helps to refine the details of small objects. After adding ECN to the baseline, both PAcc and mIoU were higher than the baseline, 83.1%, and 65.5%, respectively. The recognition accuracy of individual categories is higher than the baseline, especially "wreck," "reef," "sediment," and "water" which were much higher than the baseline. When ADC and ECN were added to the baseline at the same time, the data were significantly improved, and the PAcc and mIoU were as high as 86.9% and 68%, which increased by 7.3% and 4.2% compared to the baseline. The ADC and ECN numbers are much better than

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

Table I. *Experimental data of ADC and ECN ablation.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	PAcc	mIoU
Baseline	80.5	78.2	77.5	66.4	60.1	81.5	68.7	62.1	79.6	63.8
+ADC	82.1	83.1	77.8	68.4	62.3	82.0	72.6	64.1	82.3	65.1
+ECN	83.9	82.4	78.6	69.2	63.4	83.1	71.5	65.8	83.1	65.5
+ADC + ECN	**87.8**	**88.4**	**82.0**	**74.3**	**67.9**	**85.9**	**78.1**	**70.6**	**86.9**	**68.0**

The black bolded font in the table indicates the excellence metrics for each algorithm.

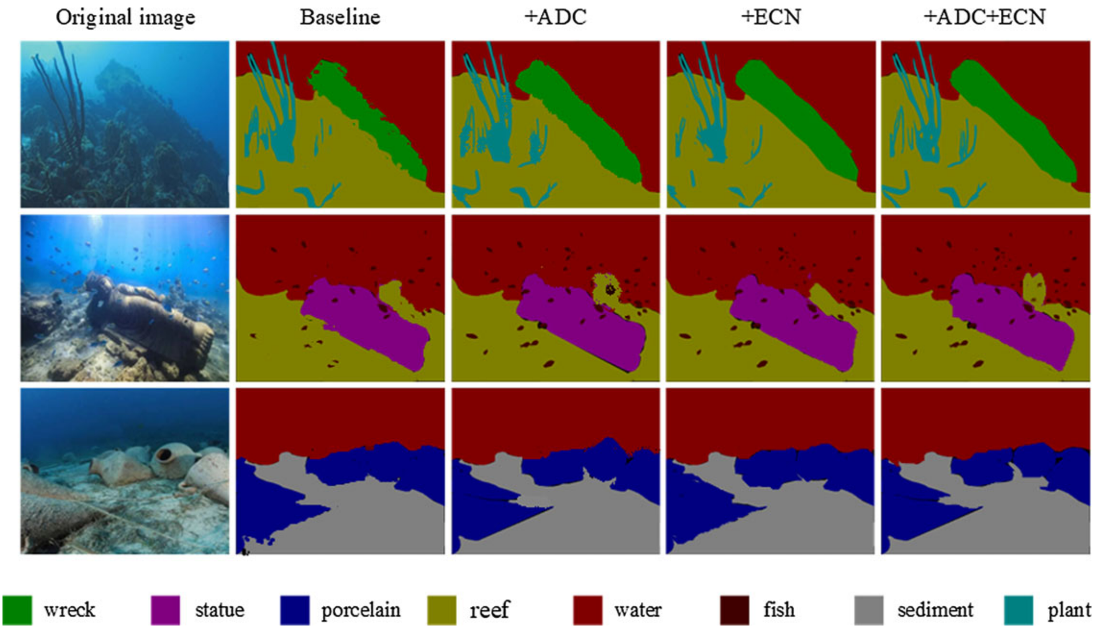


Figure 5. *Visual effect of the ablation experiment.*

adding either one of them. This is because the adaptive convolutional network can change the receptive field according to the size of the object, to extract more features. Strengthening the classification network increases the distance between similar categories and makes classification results more accurate. The combination of the two not only improves the evaluation indexes but also enables the analysis of more complex scenarios. The experimental results demonstrated the effectiveness of ADC and ECN.

From the figure, it can be seen that the baseline network is less effective in recognizing details and small objects (e.g., fish). After adding the ADC network, the overall recognition effect is significantly improved. In the second graph of the second column, it can be seen that the number of recognized fish has significantly increased contour clarity, indicating a better recognition rate for small objects. Adding the EDC network, the effect graph can see the different object boundary contours are clear and the recognition is better. By adding both ADC and EDC networks, the overall visual effect is significantly improved. The visual effect graph illustrates that both ADC and EDC networks proposed in this paper can effectively improve the algorithm's performance.

4.4. Comparison of different dilated convolutional networks

Dilated convolutional networks were proposed long ago [37], and many current studies on dilated convolution have been improved upon. In order to verify the advancedness of the proposed adaptive dilated

Table II. *Comparison results of different dilated convolutions.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	PAcc	mIoU
DCLS [19]	86.1	85.4	**82.3**	73.8	67.2	85.3	77.3	70.4	86.1	67.4
AFDC [20]	**88.1**	87.1	81.4	72.9	65.4	85.4	76.3	68.7	85.8	67.2
ADC [21]	87.9	86.9	81.9	74.1	66.7	**86.0**	75.8	69.3	86.3	67.6
Ours	87.8	**88.4**	82.0	**74.3**	**67.9**	85.9	**78.1**	**70.6**	**86.9**	**68.0**

The black bolded font in the table indicates the excellence metrics for each algorithm.

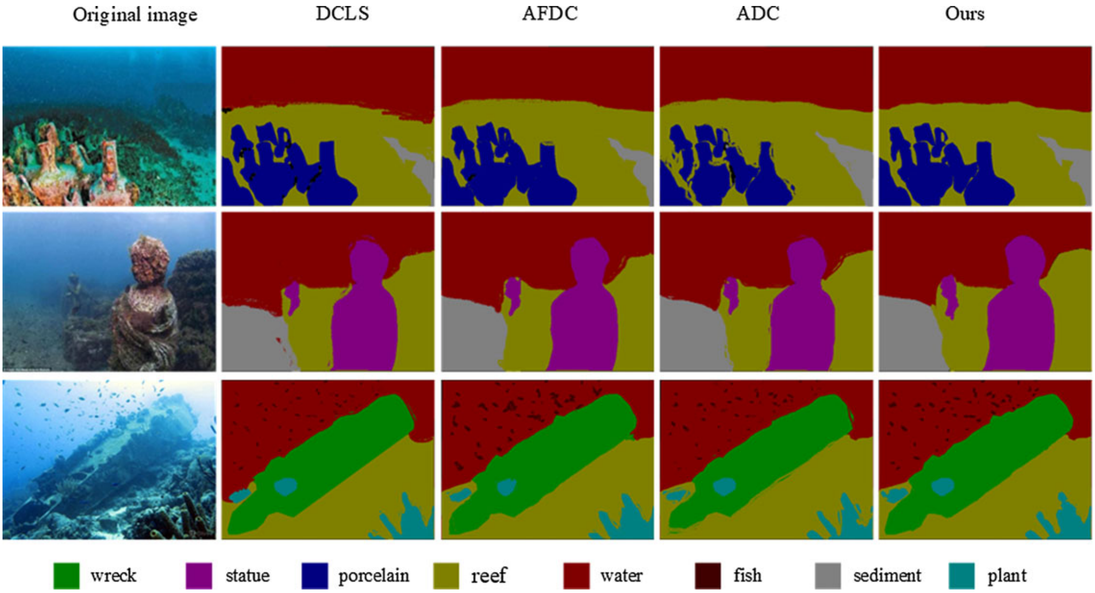


Figure 6. *Visual effect of different dilated convolution comparison.*

convolution, the proposed dilated network is replaced by the current state-of-the-art expanded convolutional network for comparison experiments. The experimental results are shown in Table II. The visual effect graph is shown in Fig. 6.

As can be seen in Table II, the adaptive dilated convolution proposed in this paper performs better in terms of performance, with the highest PAcc and mIoU. For each category, the recognition accuracy performs best overall. However, other methods also have the highest recognition accuracy for a single category. The highest accuracy of porcelain recognition by DCLS is 82.3%, which is higher than 0.3% in the dilated convolutional network proposed in this paper. The maximum pixel accuracy of AFDC for wreck is 88.1%, which is higher than the proposed dilated convolutional network by 0.4%. The highest recognition accuracy of ADC for water is 86.0%, which is 0.1% higher than the proposed algorithm. All the above methods obtain more context information by increasing receptive field, but the overall recognition accuracy in complex scenes with both large and small objects is slightly lower than that of the network proposed in this paper. The proposed adaptive dilated convolution can change the size and shape of the receptive field arbitrarily according to the size and shape of different target categories, thereby achieving better performance in PAcc and mIoU on the USS dataset.

As can be seen in Fig. 6, the visual effects of different dilated convolutions for different classes of object recognition are basically the same. However, the adaptive dilated convolution proposed in this paper outperforms the other dilated convolutions in terms of the overall object recognition contour and the recognition accuracy of small objects. From the visual results, it can be seen that DCLS performs poorly on low-resolution images because using a learnable spatial sampling rate can result in small kernel sizes that do not capture sufficient contextual information, thus affecting segmentation performance.

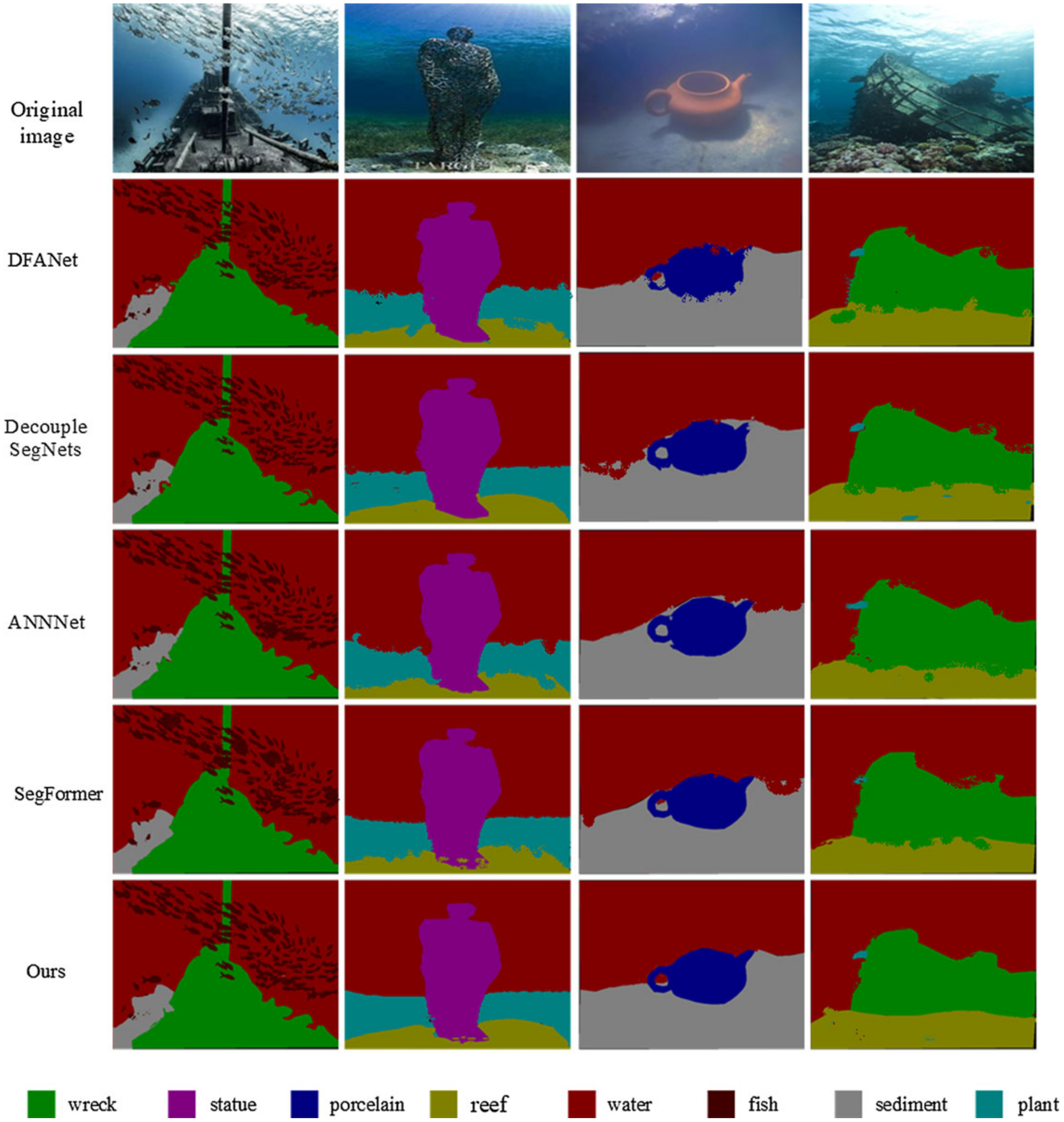


Figure 7. *General scene parsing visual effect.*

Although AFDCNet can adapt to different object sizes, its performance is not as good as other networks when dealing with dense small objects. As can be seen from the third image in the first row, when the target part in the scene is occluded or deformed, the prediction performance of ADC is significantly lower than other methods.

4.5. Experimental results and analysis under different conditions

4.5.1. The parsing results under the general condition

To evaluate the performance of the proposed algorithm in regular scenes, it is compared with the existing more advanced methods. The visual results of its detection are shown in Fig. 7, class pixel accuracy and PAcc as shown in Table III, and the class intersection-over-union and mIoU as shown in Table IV.

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

Table III. *Class pixel accuracy and PAcc under regular conditions.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	PAcc
DFANet [11]	85.9	82.0	80.5	70.6	64.8	84.9	73.6	68.2	82.1
DecoupleSegNets [12]	87.9	83.8	**83.5**	69.9	63.8	87.0	74.8	67.4	83.8
ANNNet [28]	**90.5**	87.2	82.0	69.5	64.9	**88.8**	75.4	66.0	84.4
SegFormer [30]	88.7	87.6	81.2	67.8	66.9	86.4	76.5	69.5	85.2
Ours	87.8	**88.4**	82.0	**74.3**	**67.9**	85.9	**78.1**	**70.6**	**86.9**

The black bolded font in the table indicates the excellence metrics for each algorithm.

Table IV. *Class intersection-over-union and mIoU under conventional conditions.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	mIoU
DFANet [11]	75.4	72.1	70.1	60.3	53.9	75.7	63.9	54.9	65.5
DecoupleSegNets [12]	76.1	72.9	**72.6**	61.2	53.7	76.2	64.2	54.6	66.1
ANNNet [28]	**79.8**	76.0	71.3	63.1	54.0	**77.9**	64.5	54.0	67.0
SegFormer [30]	76.4	76.2	72.0	62.8	55.1	76.1	67.5	55.8	67.1
Ours	76.2	**76.5**	71.9	**64.3**	**55.3**	75.8	**67.8**	**56.0**	**68.0**

The black bolded font in the table indicates the excellence metrics for each algorithm.

From the experimental results data in Table III and Table IV, we can see that the PAcc and mIoU of the proposed algorithm are higher than other algorithms, 86.9% and 68%, respectively. The pixel accuracy and mIoU of all eight categories are improved compared with DFANet, which has increased the perception field, but it is difficult to identify different object categories accurately because of the inconsistent size and large difference of object categories in underwater shipwreck scenes. Especially for small objects porcelain and fish, the recognition rate is the worst among all methods, 80.5% and 73.6%. Although DFANet increases the receptive field, the size of object classes in underwater shipwreck scenes is inconsistent and varies greatly. Therefore, it is difficult to recognize different object categories accurately. DecoupleSegNets uses a feature pyramid-based approach to process different scales of input images, but it is still sensitive to the input image size. The accuracy of porcelain recognition is higher than the proposed algorithm by 0.7%, while all other categories are lower than the algorithm in this paper. The attention mechanism and non-local neural network structure of ANNNet make it have certain black box nature and poor interpretation for some categories. SegFormer has a certain limitation on the size of the input image, which needs to be adjusted within a certain range. And the adaptive dilated convolution proposed in this paper can automatically adjust the size of the perceptual field according to the object category. It can fully extract multi-scale features of different target classes. The category pixel accuracy and mIoU of the proposed method are mostly higher than the other three algorithms. The algorithm proposed in this paper has the highest recognition accuracy for statue, reef, sediment, and fish under general conditions and is significantly higher than the other compared algorithms. The above comparative experimental results show that adequate feature learning, construction of contextual semantic relations, and effectively enhanced classification can significantly improve the performance of the proposed method on the USS dataset.

From the visual effect figure, it can be found that for dense and relatively small objects (fish), it is easy to recognize multiple as one. The boundary recognition result is blurred when different objects intersect. There is a phenomenon of miscalculation in the figure, for example, “reef” is recognized as “wreck.” Overall, the overall recognition degree of the proposed method for different object categories and the boundary contours are relatively good, which proves that the performance of the proposed algorithm on the USS dataset is better than other algorithms.

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

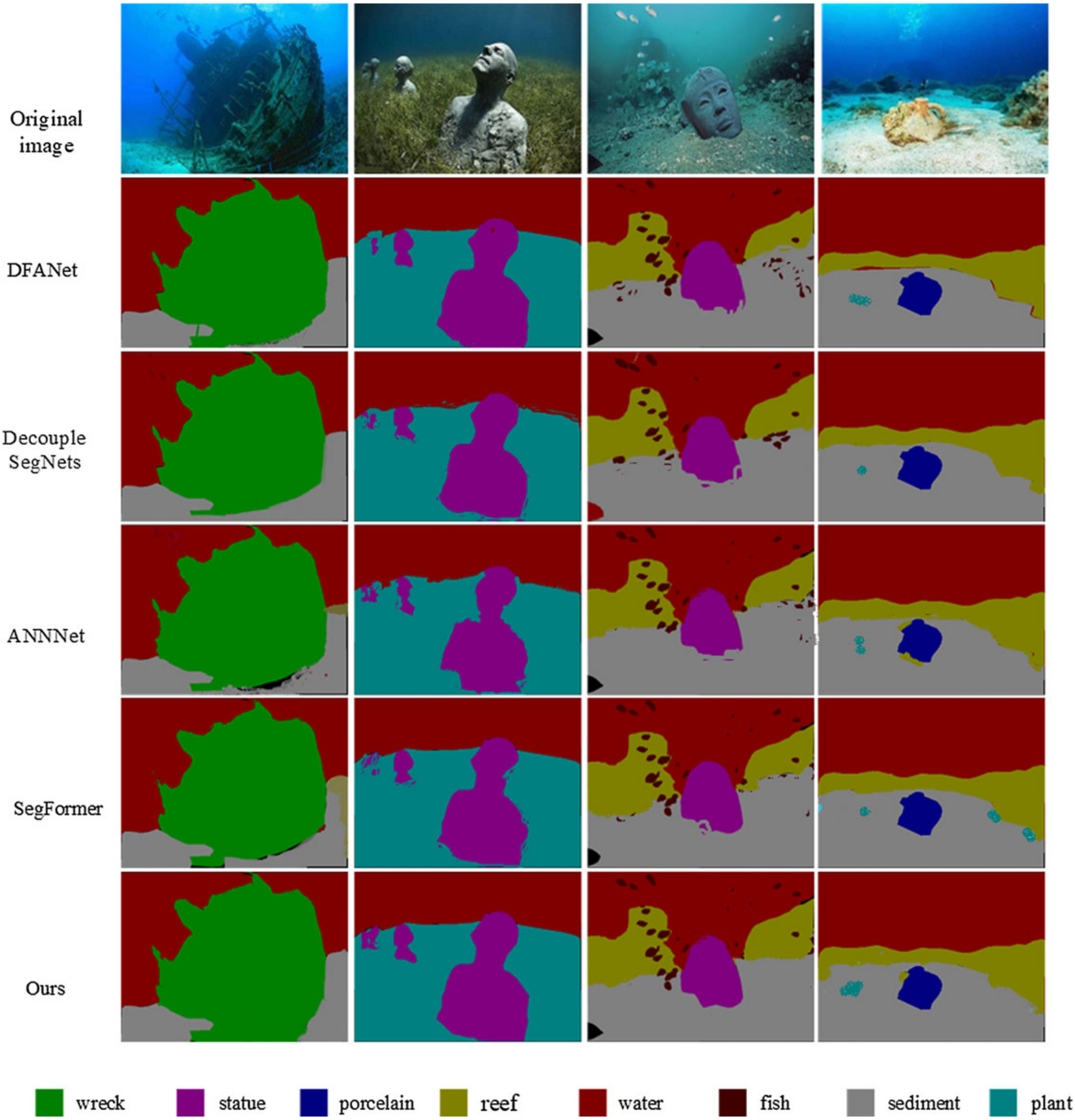


Figure 8. *Semi-burial situation of cultural relics scene parsing visual effect.*

4.5.2. The parsing results under the semi-burial situation of cultural relics

To verify the effectiveness of the proposed algorithm under different situations. The visual effects are compared with other methods in the case of semi-buried cultural relics, and the visual effects are shown in Fig. 8. Class pixel accuracy and PAcc with are shown in Table V, and class intersection-over-union and mIoU are shown in Table VI.

From the experimental results data in Table V and Table VI, it can be seen that the PAcc and mIoU of the algorithm proposed in this paper are higher than the other algorithms in the case of semi-buried cultural relics, 85.4% and 65.4%, respectively. From the data in the table, it can be seen that most of the target categories of the algorithm proposed in this paper are higher than other algorithms. The pixel accuracy and mIoU of the proposed algorithm are lower than those of the DFANet algorithm by 1.2% and 0.6%, respectively. The pixel accuracy and mIoU for fish are lower than the algorithm ANNNet by 0.6% and 0.3%, respectively. Pixel accuracy and mIoU for the plant are lower

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

Table V. *Class pixel accuracy and PAcc under semi-buried conditions for cultural relics.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	PAcc
DFANet [11]	81.3	80.2	78.3	63.7	**64.5**	83.1	73.8	68.4	82.5
DecoupleSegNets [12]	82.9	79.5	80.1	66.2	61.2	82.2	74.6	**69.2**	81.3
ANNNet [28]	83.9	82.3	79.6	65.2	60.8	84.3	**76.7**	65.8	82.7
SegFormer [30]	83.6	82.5	77.8	64.5	62.4	84.9	75.4	68.3	82.9
Ours	**84.5**	**83.1**	**81.9**	**67.8**	63.2	**85.6**	76.2	68.1	**85.4**

The black bolded font in the table indicates the excellence metrics for each algorithm.

Table VI. *Class intersection-over-union and mIoU under semi-buried conditions for cultural relics.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	mIoU
DFANet [11]	72	70.2	66.4	53.5	**53.8**	74.1	64.3	55.2	63.7
DecoupleSegNets [12]	72.9	69.9	69.7	54.1	52.4	72.2	64.6	**55.6**	63.9
ANNNet [28]	72.5	71.8	68.1	56.1	51.9	73.3	**65.2**	54.9	64.2
SegFormer [30]	72.4	72.3	67.6	55.4	52.9	74.3	64.5	55.1	64.3
Ours	**73.6**	**72.6**	**71.8**	**57.3**	53.2	**74.9**	64.9	55.2	**65.4**

The black bolded font in the table indicates the excellence metrics for each algorithm.

than the algorithm DecoupleSegNets by 1% and 0.4%, respectively. However, the pixel accuracy and mIoU of semi-buried artifacts (wreck, statue, porcelain) are significantly higher than the other algorithms. In the case of semi-buried artifacts, adaptive dilated convolution can use the dilated coefficient of the convolution kernel to expand the perceptual field, which in turn improves the feature extraction ability of the target class. At the same time, adaptive dilated convolution can also utilize the size and shape of the convolution kernel for feature enhancement and compensation to improve the expression ability and recognition performance of the target category. The comparison results of experimental data show that adjusting the perceptual field size according to the object size can fully extract object features, fuse contextual information features, and improve the segmentation accuracy of object categories.

It can be seen from the visual effect figure that the recognition of the semi-buried cultural relics algorithm proposed in this paper has the highest completeness in the recognition of cultural relics. As can be seen in the experimental effect plots of other methods, there exist buried cultural relics with plants or sediment being recognized as part of the cultural relics, resulting in blurred edges of the object class. For some smaller object categories, there also exists the phenomenon that they cannot be recognized. The comprehensive experimental effect figure can be found that the proposed method in the paper has a high degree of completeness in artifact recognition and a relatively good edge recognition effect, which further proves the effectiveness of the proposed algorithm.

4.5.3. The parsing results under the turbid water condition

Due to the presence of a large amount of sediment in the site where the cultural relics are located, the water quality of the environment in which they are located is turbid. In order to further verify the effectiveness of the proposed algorithm, the visual effect diagram is shown in Fig. 9 for comparison with other algorithms in the case of turbid water, the category detection accuracy and detection pixel accuracy are shown in Table VII, the category cross-accuracy and mIoU are shown in Table VIII.

From the experimental result data in Table VII and Table VIII, we can see that the PAcc and mIoU of the proposed algorithm in this paper are higher than other algorithms, which are 84.9% and 65.5%, respectively. Meanwhile, the category pixel accuracy and mIoU of eight categories are improved, which are higher than other algorithms. Compared with DFANet, the proposed algorithm improves pixel accuracy by 3.4% and mIoU by 1.3%. The pixel accuracy of the algorithm presented in this paper is 3% and 1.2% higher than that of DecoupleSegNets. The pixel accuracy and mIoU of ANNNet are 2.9%

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

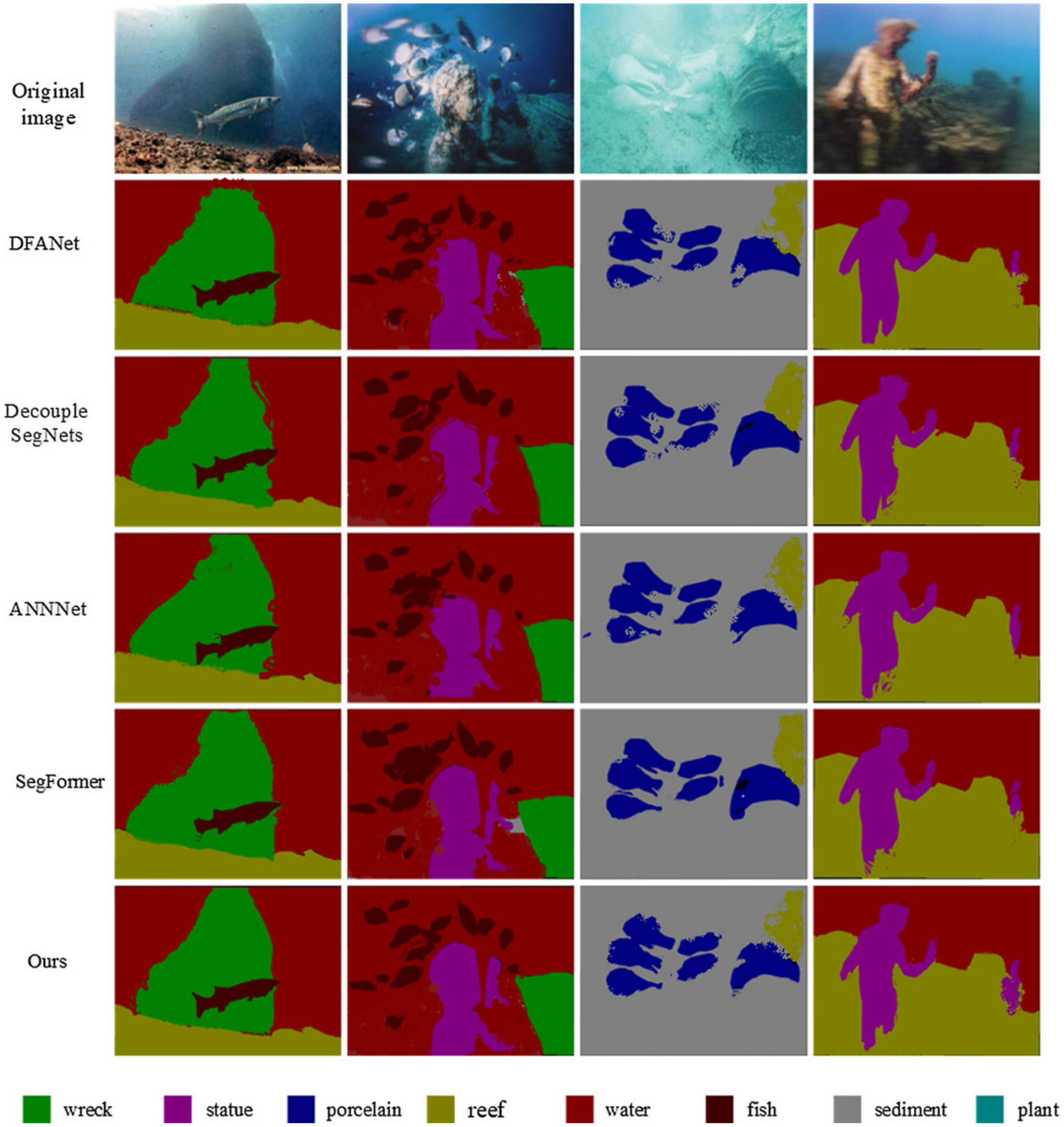


Figure 9. *Visual effect of scene parsing under turbidity condition.*

Table VII. *Class pixel accuracy and pixel accuracy under turbid water conditions.*

Model	**Wreck**	**Statue**	**Porcelain**	**Reef**	**Sediment**	**Water**	**Fish**	**Plant**	**PAcc**
DFANet [11]	82.8	81.1	78.3	68.5	63.4	82.5	72.9	67.8	81.5
DecoupleSegNets [12]	83.5	80.5	82.1	67.4	62.8	81.8	72.3	66.9	81.9
ANNNet [28]	82.7	83	80.7	68.2	65.1	83.2	73.4	65.4	82
SegFormer [30]	83.9	82.7	78.6	65.2	66.9	82.1	74.8	68.7	82.9
Ours	**84.2**	**83.5**	**82.9**	**71.9**	**67.2**	**83.6**	**75.9**	**69.3**	**84.9**

The black bolded font in the table indicates the excellence metrics for each algorithm.

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

Table VIII. *Class intersection-over-union and mIoU under turbid water conditions.*

Model	Wreck	Statue	Porcelain	Reef	Sediment	Water	Fish	Plant	mIoU
DFANet [11]	72	71.9	67.9	57.9	53.2	72.4	63.6	54.7	64.2
DecoupleSegNets [12]	72.8	71.4	69.8	57.1	53	71.3	63.3	55.3	64.3
ANNNet [28]	72.1	72.9	71.3	56.4	53.6	72.4	63.5	53.8	64.5
SegFormer [30]	73.1	72.4	70.6	55.7	53.9	72.1	64.1	55.3	64.7
Ours	**73.3**	**73.1**	**72.2**	**58.2**	**54.2**	**72.9**	**64.3**	**55.4**	**65.5**

The black bolded font in the table indicates the excellence metrics for each algorithm.

and 1% lower than that of the proposed algorithm. Compared with SegFormer, the pixel accuracy of the algorithm proposed in this paper is improved by 2% and that of mIoU by 0.8%. When there are ambiguous and difficult-to-judge samples, a single classifier may not be able to accurately classify these samples because they may belong to multiple classes or not to any one class. However, stepwise classification using multiple classifiers can better handle these ambiguous samples. In a multi-classifier model, each classifier is responsible for classifying a set of categories. First, the initial classifier classifies all samples, and misclassified or ambiguous samples, they are sent to the next classifier for further classification. This process can be repeated several times until all samples are assigned to the last classifier for final classification. The experimental results show that the proposed algorithm can effectively improve the segmentation performance of object categories under the turbid water condition, which verifies the effectiveness of the proposed network.

As can be seen from the visual effect diagram, all the category boundaries are blurred, which is caused by the blurring of the extracted features due to the turbidity of the water, which makes it difficult to discriminate in the classification process. And it can be seen from the figure that there is the phenomenon of object category misclassification. Due to the blurred water quality, the features are difficult to extract, and some objects are identified as similar backgrounds. For example, “porcelain” is identified as “sediment,” and “statue” is identified as “reef.” From the last line of the effect, we can see that our proposed method has a good visual effect, which proves the effectiveness of the proposed method in turbid water conditions.

4.6. Comparative experiments on public datasets

To verify the generalization of the algorithm proposed in this paper, we conducted comparison experiments with current state-of-the-art algorithms on a public dataset containing the ADE20K dataset, the Cityscapes dataset, and the SUIM dataset. ADE20K covers various annotations of scenes, objects, and object parts, and contains various objects in natural space environments. Each image has an average of 19.5 instances and 10.5 object classes with 150 semantic categories. The Cityscapes dataset contains 20,210 training images, 2000 validation images, and 3351 test images. 30 different categories of city street data from 50 different cities for different periods and seasons are included in the Cityscapes dataset. The SUIM dataset is the first publicly available large-scale underwater semantic segmentation dataset with 2975 training images, 1525 test images, and 500 validation images. The dataset has 1525 images and contains 8 categories: Background (waterbody) (BW), Human divers (HD), Aquatic plants and seagrass (PF), Wrecks or ruins (WR), Robots (RO), Reefs and invertebrates (RI), Fish and vertebrates (FV), and Sea-floor and rocks (SR).

4.6.1. ADE20K and Cityscapes comparison test

Since the experimental results on the ADE20K and Cityscapes datasets are similar, we put the two experimental results together for analysis. Due to the large number of scene categories in these two datasets only PAcc and mIoU are reported in this paper, and the experimental data are shown in Tables IX and X, and the visual effect graphs of scene resolution are shown in Figs. 10 and 11. The experimental data

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

Table IX. *PAcc and mIoU of the ADE20K dataset.*

Model	**PAcc**	**mIoU**
DFANet [11]	80.8	43.6
DecoupleSegNets [12]	81.5	43.9
ANNNet [28]	79.4	42.4
SegFormer [30]	82.8	44.8
Ours	**82.9**	**45.1**

The black bolded font in the table indicates the excellence metrics for each algorithm.

Table X. *PAcc and mIoU of the Cityscapes dataset.*

Model	**PAcc**	**mIoU**
DFANet [11]	94.6	81.5
DecoupleSegNets [12]	96.2	82.8
ANNNet [28]	95.4	82.3
SegFormer [30]	96.9	83.2
Ours	**97.5**	**83.4**

The black bolded font in the table indicates the excellence metrics for each algorithm.

in Table IX shows that the proposed algorithm achieves the highest PAcc and mIoU on both ADE20K dataset and the Cityscapes dataset. DFAT is not fine enough for object edge detail processing, resulting in lower data and a less accurate visual effect of scene parsing results. The attention mechanism in ANNNet is based on a homogeneous image grid, which is less effective in resolving irregular shapes in the scene. SegFormer performs differently on different size input images. The overall effect of the proposed algorithm and the current state-of-the-art algorithm on visual scene resolution is the same, but the proposed algorithm performs better in detail. The proposed algorithm can make a consistent prediction for large objects and an accurate prediction for small objects by expanding adaptive convolution.

4.6.2. SUIM dataset comparison test

To further verify the advancedness and performance of the algorithm proposed in this paper for underwater images, we conducted comparison experiments between current advanced algorithms and the algorithm proposed in this paper on the SUIM dataset. In addition, we added the SUIMNet method to this set of comparison experiments. The experimental data are shown in Table XI, and the visual effect graph of scene resolution is shown in Fig. 12.

The data in the table show that the proposed algorithm in this paper has the highest PAcc and mIoU among the comparison methods, which are 93.1% and 81.1%, respectively. In terms of category pixel accuracy, the algorithm proposed in this paper is slightly lower than other algorithms for the whole part of the category. In terms of category pixel accuracy, SUIMNet has the highest accuracy for BW and FV pixels, which is 0.2% and 0.8% higher than the proposed algorithm. decoupleSegNets has the best recognition for HD with 90.2%, which is 0.3% higher than the proposed algorithm. ANNNet has the highest recognition accuracy for WR with 86.3%, which is 0.5% higher than the proposed algorithm. The visual resolution of this paper is clearer than that of this paper, as can be seen from the visual effect graph. This is because the SUIM dataset is larger and can provide more samples to train the model, thus better capturing the statistical patterns of the data. In addition, the SUIM dataset already contains many high-quality underwater images with clearer category features than the USS dataset. The higher-clarity images are easier to be classified during the training process because the object boundaries and details are clearer.

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

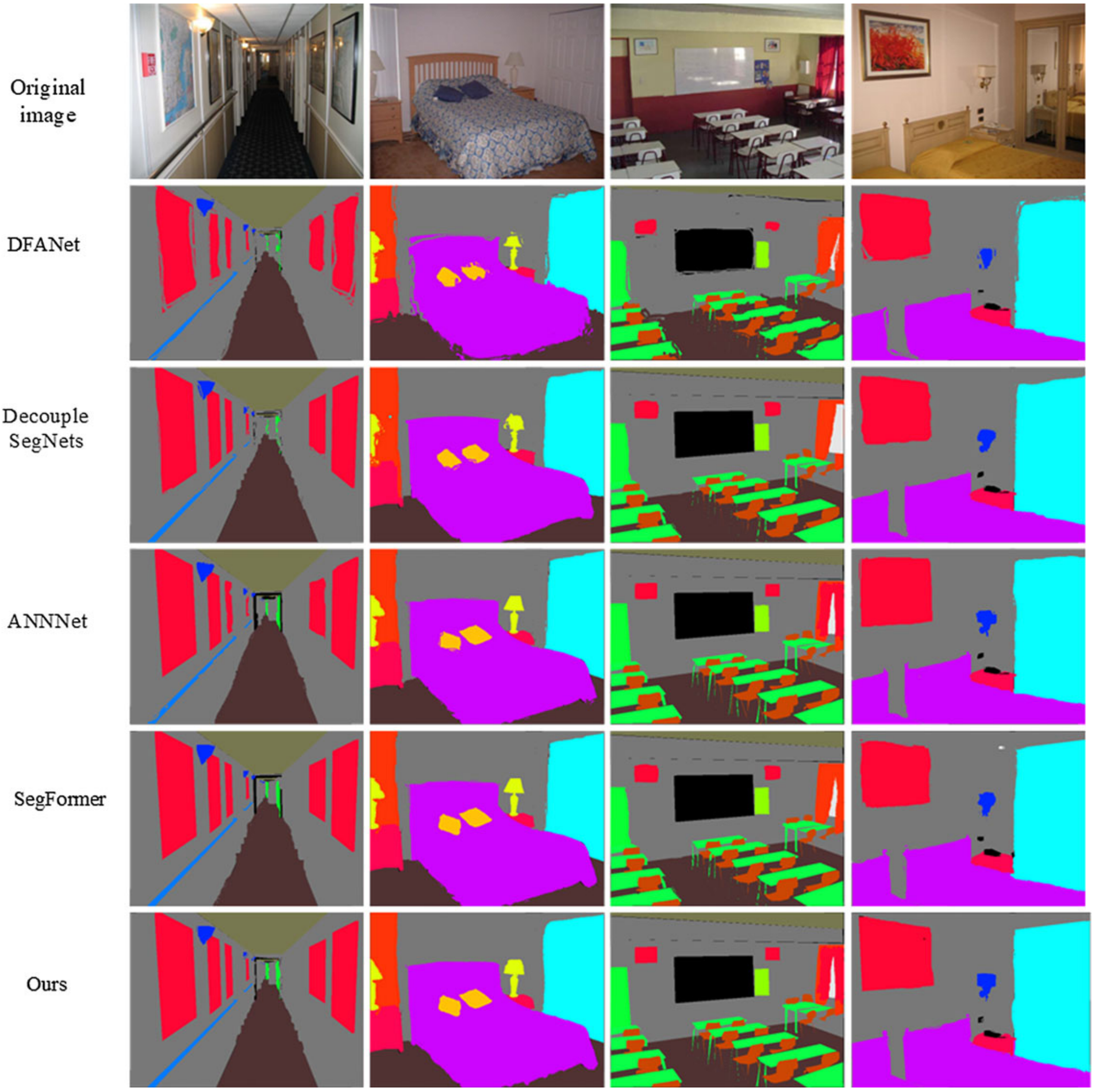


Figure 10. *ADE20K dataset scene parsing effect.*

Table XI. *PAcc and mIoU of the SUIM dataset.*

Model	BW	HD	PF	WR	RO	RI	FV	SR	PAcc	mIoU
DFANet [11]	96.3	89.1	80.1	83.1	74.3	81.2	90.4	72.4	91.1	80.1
DecoupleSegNets [12]	95.4	**90.2**	79.2	81.2	73.1	83.2	91.5	73.8	90.2	79.7
ANNNet [28]	96.2	89.6	82.8	**86.3**	75.2	81.7	92.3	74.2	91.5	80.2
SegFormer [30]	96.8	87.8	83.1	79.8	76.4	84.2	91.9	71.9	91.9	80.3
SUIMNet [33]	**97.5**	89.9	82.9	86.2	77.8	84.5	93.7	75.3	92.3	80.5
Ours	97.3	88.9	**83.3**	85.8	**78.2**	**84.7**	92.9	**75.7**	**93.1**	**81.1**

The black bolded font in the table indicates the excellence metrics for each algorithm.

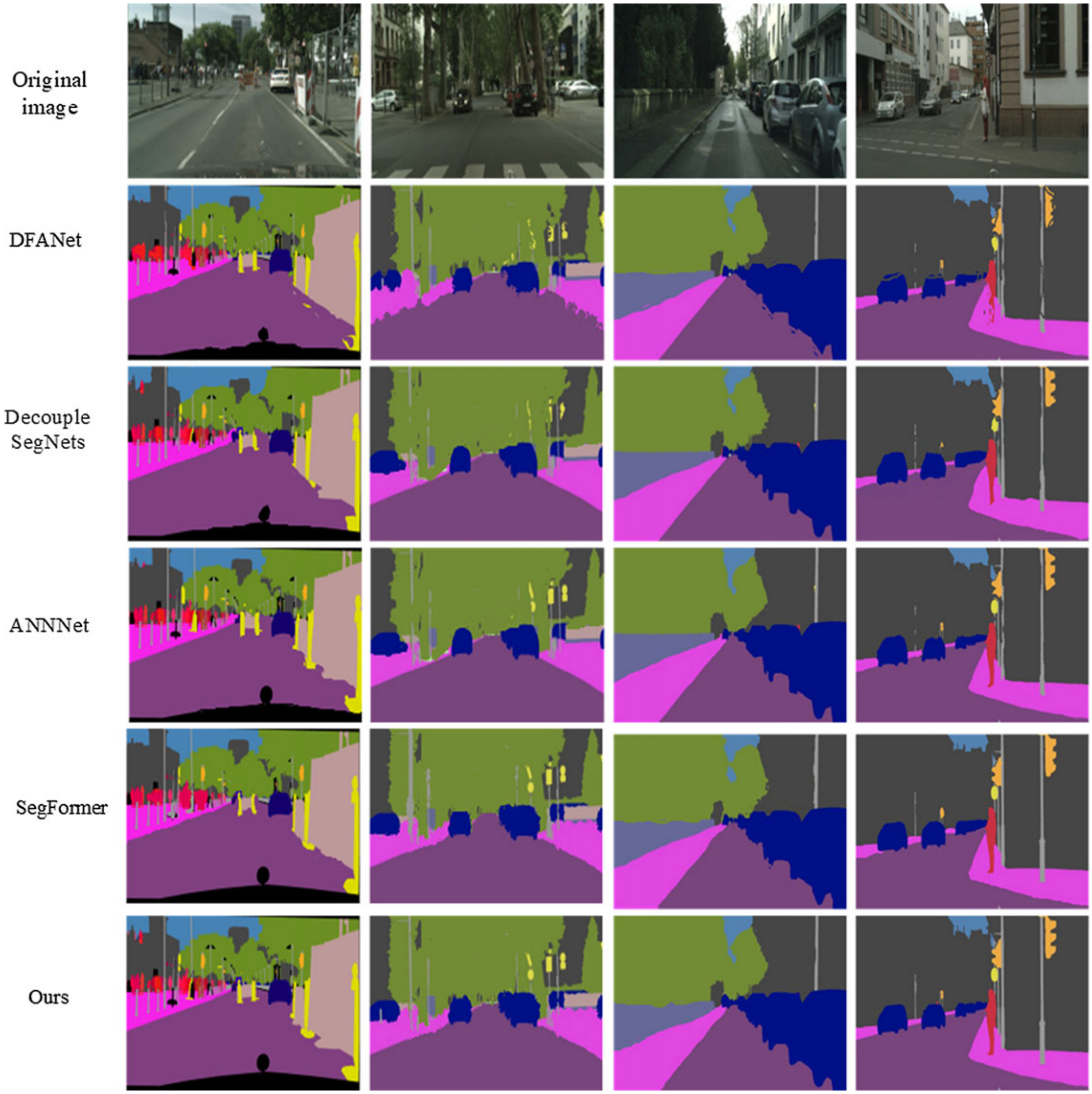


Figure 11. *Cityscapes dataset scene parsing effect.*

5. Conclusions

In this paper, we propose a GENet scene parsing method for underwater archaeological scenarios to solve the problems encountered in this scenario. Adaptive dilated convolution is proposed to obtain flexible small and large and shaped receptive fields for scene resolution. The adaptive dilated convolution reshapes the relevant receptive fields by learning the vector of dilated coefficients to adaptively change the convolutional ground patches and shapes. In addition, we propose a reinforced classification network, which optimizes the classifier based on the difference-ground regularization method and can effectively distinguish confusable categories. Besides, a self-made USS dataset is developed for underwater archaeological scenes. And extensive experiments are conducted on the USS dataset to verify the effectiveness of the proposed method. The effectiveness of the proposed network is demonstrated by ablation experiments. The experimental results in comparison with the current state-of-the-art dilated convolutional network show that the adaptive expanded convolution has excellent performance. Finally, we compare with the current state-of-the-art methods in three different cases, and both PAcc and mIoU

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

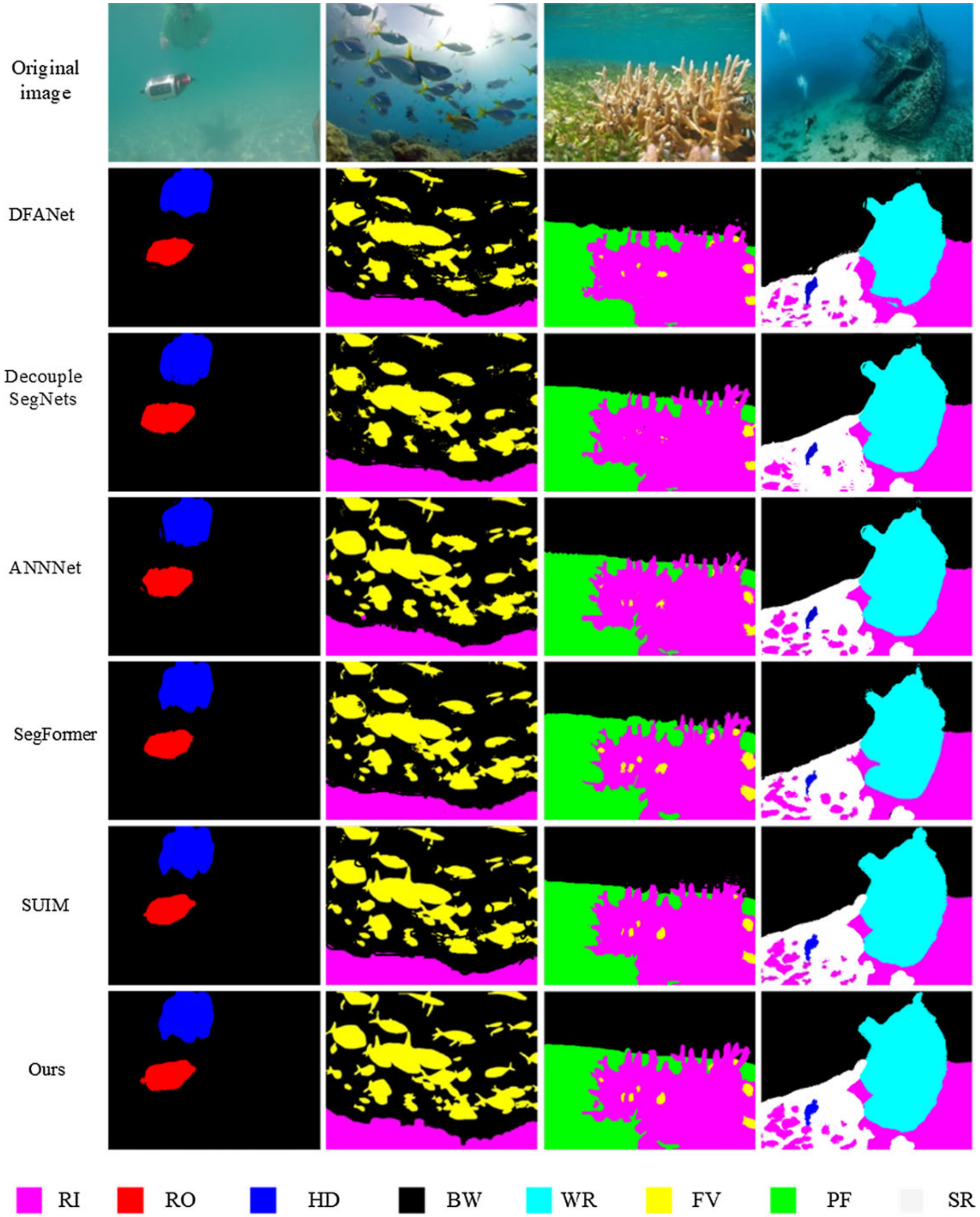


Figure 12. *SUIM dataset scene parsing effect.*

are higher than other algorithms, proving the superiority of our algorithm. To verify the generalization of the algorithm proposed in this paper, we did comparison experiments on the public dataset ADE20K, cityscapes, and underwater dataset SUIM, and the experimental results show that the algorithm proposed in this paper performs well on the public dataset and outperforms other algorithms. In the comparative experiments on public datasets, we can see that the experimental results of the same algorithm on public

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

datasets are significantly higher than the homemade datasets in this paper. This is because the public dataset has various scenes and complex and changeable categories. The adaptive expansive convolution convex polygons with the arbitrary variation of receptive field proposed in this paper apply to images with certain rules, but the effect is not satisfactory for complex image structures and irregular edges. In addition, the USS dataset proposed in this paper has certain limitations. The dataset contains few scene and object categories and a small amount of data, so the model cannot be fully trained. Our next step will be to optimize the performance of the proposed algorithm and expand the size of the data set to improve the diversity of the scene data. The proposed method and USS dataset can be applied more widely.

Authors' contributions. Junyan Pan wrote the article. Jishen Jia reviewed and edited this article. Lei Cai conceived and designed the study.

Financial support. This work was supported by the Science and Technology Project of Henan Province (grant number 222102110194, 222102320380, 232102320338).

Competing interests. The authors declare that there are no Competing interests regarding the publication of this article.

Ethical considerations. None.

References

[1] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," **In:** *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA (2015) pp. 3431–3440.

[2] J. Jiang, J. Liu, J. Fu, X. Zhu, Z. Li and H. Lu, "Global-guided selective context network for scene parsing," *IEEE Trans. Neural Netw. Learn. Syst.* **33**(4), 1752–1764 (2022).

[3] W. Zhou, X. Lin, J. Lei, L. Yu and J. Hwang, "MFFENet: Multiscale feature fusion and enhancement network for RGB-thermal urban road scene parsing," *IEEE Trans. Multimedia* **24**(6), 2526–2538 (2021).

[4] S. Ma, Y. Pang, J. Pan and L. Shao, "Preserving details in semantics-aware context for scene parsing," *Sci. China Inf. Sci.* **63**(2), 1–14 (2020).

[5] K. Yan, H. Wang, S. Bu, L. Yang and J. Li, "Scene parsing for very high resolution remote sensing images using on attention-residual block-embedded adversarial networks," *Remote Sens. Lett.* **12**(7), 625–635 (2021).

[6] S. Liu, H. Zang, S. Li and J. Yang, "Built-in depth-semantic coupled encoding for scene arsing, vehicle detection, and road segmentation," *IEEE Trans. Intell. Transp. Syst.* **22**(9), 5520–5534 (2021).

[7] J. Jiang, Z. He, S. Zhang, X. Zhao and J. Tan, "Learning to transfer focus of graph neural network for scene graph parsing," *Pattern Recognit.* **112**(4), 107707 (2020).

[8] Z. Xiong, Y. Yuan, N. Guo and Q. Wang, "Variational Context-Deformable ConvNets for Indoor Scene Parsing," **In:** *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA (2020) pp. 3991–4001.

[9] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang and J. Zhang, "Total 3D Understanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image," **In:** *33st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, MA (2020) pp. 52–61.

[10] J. Ji, X. C. Lu, M. Luo, M. Yin, Q. Miao and X. Liu, "Parallel fully convolutional network for semantic segmentation," *IEEE Access* **9**(11), 673–682 (2021).

[11] H. Li, P. Xiong, H. Fan and J. Sun, "DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation," **In:** *33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA (2019) pp. 9522–9531.

[12] X. Li, L. Zhang, G. Cheng, Z. Lin, S. Tan and Y. Tong, "Improving Semantic Segmentation via Decoupled Body and Edge Supervision," **In:** *European Conference on Computer Vision (ECCV)* (Springer, Cham, 2020) pp. 435–452.

[13] X. Zhang, Y. Yan, J.-H. Xue, Y. Hua and H. Wang, "Semantic-aware occlusion-robust network for occluded person re-identification," *IEEE Trans. Circuits Syst. Video Technol.* **31**(7), 2764–2778 (2021).

[14] L. Liao, J. Xiao, Z. Wang, C.-W. Lin and S. Satoh, "Uncertainty-aware semantic guidance and estimation for image inpainting," *IEEE J. Sel. Top. Signal Process.* **15**(2), 310–323 (2021).

[15] L. Cai, X. C. Qin and T. Xu, "EHDC: Enhanced dilated convolution framework for underwater blurred target recognition," *Robotica* **41**(3), 900–911 (2022).

[16] X. Qiao, Q. Zheng, Y. Cao and R. W. H. Lau, "Object-level scene context prediction," *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(9), 5280–5292 (2021).

[17] Z. Li, Y. Sun, L. Zhang and J. Tang, "CTNet: Context-based tandem network for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(12), 9904–9917 (2021).

[18] Y. Sun and Z. Li, "SSA: Semantic structure aware inference for weakly pixel-wise dense predictions without cost," *arXiv preprint* arXiv:2111.03392 (2021).

https://doi.org/10.1017/S026357472300098X Published online by Cambridge University Press

[19] I. Khalfaoui-Hassani, T. Pellegrini and T. Masquelier, "Dilated convolution with learnable spacings," *arXiv preprint* arXiv:2112.03740 (2021).

[20] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng and J. Fan, "Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment," **In:** *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA (2020) pp. 14114–14123.

[21] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan and E. Zhou, "Adaptive dilated convolution for human pose estimation," *arXiv preprint* arXiv:2107.10477 (2021).

[22] X. Niu, B. Yan, W. Tan and J. Wang, "Graphs, convolutions, and neural networks: From graph filters to graph neural networks," *IEEE Signal Process. Mag.* **37**(6), 128–138 (2020).

[23] L. Cai, C. Chen and H. Chai, "Underwater distortion target recognition network (UDTRNet) via enhanced image features," *Comput. Intell. Neurosci.* **2021**(10), 1–10 (2021).

[24] L. Cai, C. Chen, Q. Sun and H. Chai, "Glass refraction distortion object detection via abstract features," *Comput. Intell. Neurosci.* **2022**(3), 5456818 (2022).

[25] R. Liu, Z. Jiang, S. Yang and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Trans. Image Process.* **31**(7), 4922–1936 (2022).

[26] S. K. S. Rajan and N. Damodaran, "Multiscale decomposition and fusion-based color contrast restoration for various water-colored environments," *Color Res. Appl.* **47**(2), 301–328 (2022).

[27] Z. Wang, S. Zhang, W. Huang, J. Guo and L. Zeng, "Sonar image target detection based on adaptive global feature enhancement network," *IEEE Sens. J.* **22**(2), 1509–1530 (2021).

[28] Z. Zhu, M. Xu, S. Bai, T. Huang and X. Bai, "Asymmetric Non-local Neural Networks for Semantic Segmentation," **In:** *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South) (2019) pp. 593–602.

[29] W. Zhou, J. Jin, J. Lei and J. N. Hwang, "CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **60**(9), 1–10 (2021).

[30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," arXiv preprint arxiv:2105.15203 (2021).

[31] Y. Sun, Q. Chen, X. He, J. Wang, H. Feng, J. Han, E. Ding, J. Cheng, Z. Li, J. Wang, "Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning," *arXiv preprint* arXiv:2206.06122 (2022).

[32] Z. Li, H. Tang, Z. Peng, G. Qi and J. Tang, "Knowledge-guided semantic transfer network for few-shot image recognition," *IEEE Trans. Neural. Netw. Learn. Syst.* **34**(2), 1–15 (2023).

[33] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enanand and J. Sattar, "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," **In:** *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA (2020) pp. 1769–1776.

[34] S. Wang and Y. Yang, "Image semantic segmentation method based on deep fusion network and conditional random field," *Comput. Intell. Neurosci.* **2022**(5), 8961456 (2022).

[35] W. Yang and Y. Hui, "Image scene analysis based on improved FCN model," *Int. J. Pattern Recognit. Artif. Intell.* **35**(15), 2152020 (2021).

[36] P. Liu and Y. Song, "Segmentation of sonar imagery using convolutional neural networks and Markov random field," *Multidimens. Syst. Signal Process.* **31**(1), 21–47 (2019).

[37] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint* arXiv:1511.07122 (2015)

Cite this article: J. Pan, J. Jia and L. Cai (2023). "Global enhancement network underwater archaeology scene parsing method", Robotica **41**, 3541–3564. https://doi.org/10.1017/S026357472300098X