Fine reconstruction of underwater images for environmental feature fusion

Tao Xu1, Jiyong Zhou2, Wentao Guo3, Lei Cai1 and Yukun Ma1

Abstract
Complicated underwater environments, such as occlusion by foreign objects and dim light, causes serious loss of underwater targets feature. Furthermore, underwater ripples cause target deformation, which greatly increases the difficulty of feature extraction. Then, existing image reconstruction models cannot effectively achieve target reconstruction due to insufficient underwater target features, and there is a blurred texture in the reconstructed area. To solve the above problems, a fine reconstruction of underwater images with the target feature missing from the environment feature was proposed. Firstly, the salient features of underwater images are obtained in terms of positive and negative sample learning. Secondly, a layered environmental attention mechanism is proposed to retrieve the relevant local and global features in the context. Finally, a coarse-to-fine image reconstruction model, with gradient penalty constraints, is constructed to obtain the fine restoration results. Contrast experiment between the proposed algorithm and the existing image reconstruction methods has been done in stereo quantitative underwater image data set, real-world underwater image enhancement data set, and underwater image data set, clearly proving that the proposed one is more effective and superior.

Keywords
Contrastive learning, environmental attention mechanism, underwater image, fine reconstruction

Date received: 16 March 2021; accepted: 29 July 2021

Introduction
Underwater environment surveys and intelligent operations are mainly relying on underwater autonomous underwater vehicle (AUV). AUV perceives the information of surrounding environment via the visual perception system, providing important guarantee for the realization of autonomous navigation capabilities, target recognition and tracking, autonomous route planning, and coordinated control. However, there are a large number of algae, plankton, and impurities in the ocean. And the water body has a strong attenuation effect on natural light. These factors seriously affect the underwater imaging effect, resulting in occlusion of underwater images and dim light, and a large amount of target information in underwater images is missing. At the same time, the features of underwater images are difficult to extract because of the complex background information. These factors seriously affect the

1 School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang, China
2 School of Mechatronics Electrical Engineering, Henan Institute of Science and Technology, Xinxiang, China
3 The First Standardization Laboratory, China North Standardization Center, Beijing, China

Corresponding author:
Tao Xu, School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang 453003, China.
Email: xutao@hist.edu.cn
effect of underwater imaging by covering targets or attenuating light, so a large amount of target information in underwater images has been missed. Therefore, with accurate reconstruction of the images mentioned above, AUV will perform well underwater.⁴,⁵

Image reconstruction is realized by using the adjacent feature information of the missing part or the overall structure information of the image and by adopting a certain reconstruction technology to reconstruct the missing area of the image.⁶ The core technology of image reconstruction is that the global semantic structure should be maintained and vivid texture details should be generated as well in the image reconstruction area.⁷–⁹ Most of the traditional reconstruction methods are based on texture synthesis and structure-based reconstruction methods to achieve image reconstruction. The reconstruction methods based on texture synthesis all use low-level features or patches to reconstruct the missing areas of the image.¹⁰–¹⁴ The structure-based reconstruction method is based on the structure of the image and reconstructs the missing area through the method of gradual diffusion.¹⁵–¹⁷ The traditional reconstruction methods are only fit for the scenes with single background or small missing area, exposing significant deficiency for the underwater images in complex scenes, or with a large amount of missing information. This type of image requires a high degree of understanding of information to achieve underwater image reconstruction.

The rapid development of deep learning methods has provided a new avenue for image reconstruction.¹⁸–²⁰ The image reconstruction method based on deep learning heavily trains the data of the database in a deep learning network, enabling the reconstruction model to learn more deep-level feature information of the image. As generative adversarial networks (GANs), which are regarded as an unsupervised deep learning model, are applied in the field of image reconstruction,²¹–²³ the image reconstruction has got further development. The image reconstruction is accomplished by the structure of encoder and evaluated the authenticity of the restored image by a discriminator, so it can ensure the quality of the reconstructed image.

There are a large number of algae, plankton, and impurities in the ocean. Moreover, the visibility is very low in deep water for a strong attenuation of light from the water body. Complicated underwater environments, such as occlusion by foreign objects and dim light, can cause serious loss of underwater targets feature. Many factors cause the lack of underwater image feature information, so that the existing image reconstruction model cannot achieve effective reconstruction of these images. To solve the above problems, this article proposes a fine reconstruction algorithm of the underwater image with the target feature missing from the environment feature. Firstly, the environmental attention mechanism is proposed to retrieve relevant local and global environmental features in the context. The model makes greater use of environmental information to solve the problem of insufficient target information. Secondly, it constructs the image reconstruction network from rough to fine. This article embeds the environmental feature attention mechanism layer into the coarse reconstruction network. At the same time, the relevant feature coherent layer is embedded in the fine-grained reconstructed network to obtain more ideal reconstruction results. Finally, through the learning of positive and negative samples in image reconstruction model training, the reconstruction model pays more attention to salient features in the process of extracting features. This can improve the efficiency of training and solve the problem that the target is difficult to extract. The reconstruction results of the model in this paper are shown in Figure 1. The principle is shown in Figure 2. Successful reconstruction of underwater images in the presence of occlusions will not only facilitate autonomous control and collaborative mission planning of AUVs but also have a significant impact on ocean exploration, underwater research, and underwater operations. The contribution of this article can be summarized in three points:

1. Construct an image reconstruction network from rough to fine. And embedding relevant feature coherence layers in the fine reconstruction network for further refinement and perfection of the rough reconstruction results. This image reconstruction model is applied to the field of underwater

Figure 1. The reconstruction results of reconstruction model in this article. The top two images are the input underwater images with occlusion phenomenon. The lower two images are the results of the reconstruction obtained by our reconstruction model.
incomplete images. It will be helpful for underwater environment survey and underwater intelligent operation.

2. Proposes a hierarchical environmental attention mechanism, which can make greater use of background information. Solving the problem of ineffective image reconstruction due to insufficient data.

3. For the first time, the idea of positive and negative sample learning is applied to the field of underwater images. It can improve the efficiency of training and solve the problem of difficult target extraction. The learning and reconstruction capabilities of the image reconstruction model have been improved.

The rest of this article is shown as follows: the second section introduces related work. The third section introduces the principle of underwater image reconstruction algorithm proposed in this article. The fourth section introduces the design simulation based on the proposed algorithm and analyzes the experimental results. The fifth section draws conclusions and points out future research work.

**Related work**

*Image reconstruction*

The current image reconstruction methods are mainly divided into two categories. One is a nonlearning image reconstruction method and the other is an image reconstruction method based on learning. Nonlearning image reconstruction methods mainly achieve image reconstruction of missing area by diffusing neighboring information or copying information from the most relevant area of the background.\textsuperscript{10,15} This approach produces smooth and realistic results; however, its computational cost and memory usage are very large. To solve this problem, Barnes et al. proposed a fast nearest-neighbor calculation method for image reconstruction,\textsuperscript{24} which greatly improved the calculation speed and can obtain a high-quality reconstruction effect. Nonlearning methods are very effective for surface texture synthesis. But they cannot generate semantically meaningful content. So the nonlearning image reconstruction method is not suitable for processing large missing areas.

As for learning-based image reconstruction methods, incomplete image reconstruction is usually realized by making use of deep learning and GAN strategies. Pathak et al.\textsuperscript{25} proposed an image reconstruction method based on upper and lower feature prediction, which extracts the depth features of the entire image through a context encoder (CE) and makes reasonable assumptions about the generation of missing parts. At the same time, the pixel loss function is used to make the generated result clearer. However, it does not work well in fine texture generation. To solve this problem, Chao et al. proposed a multiscale image reconstruction model based on the combination of image features and texture constraints.\textsuperscript{26} This method synthesizes lifelike high-frequency details through the encoder–decoder structure in the convolutional neural network model and the generated patches, so as to obtain a clearer and more coherent reconstruction effect. Jiahui et al. proposed a repair model based on unified prefeedback.\textsuperscript{27} The model includes a context attention layer. Its principle is that to use the feature information of the known patch is used as a convolution processor. The patch generated by the convolutional design is matched with the known contextual patch, and the patch features are weighted by softmax and deconvolved to reconstruct the generated patch using the known contextual patch to obtain a fine texture reconstruction effect. However, this method ignores the semantic relevance and feature continuity of the incomplete region. To solve this problem, Liu et al. proposed an image reconstruction model based on refined deep learning.\textsuperscript{28} This method proposes to adopt a new coherent semantic layer and retain the upper and lower semantic context structure through the coherent semantic layer, and the inferred part is more reasonable.

*Contextual attention*

The attention mechanism can retrieve image background feature information, which improves image processing capability, so it has been widely used in the field of
To improve the classification ability of images, Mnih et al. introduced the attention mechanism into the CNN, which greatly improved the learning ability of the model. Xu et al. divided the attention mechanism into two types, namely soft attention and hard attention. The relevant global features are extracted via soft attention. Extract the relevant features of small areas via hard attention. Similarly, the attention mechanism is introduced into the field of image reconstruction. More potential feature information is obtained by means of the contextual attention (CA) mechanism to improve the effect of image reconstruction.

Contrastive learning

Contrastive learning, a self-supervised learning method, principle is that the model is able to note the salient features of the samples by positive and negative sample learning, thus greatly reducing the computational effort of model training. Bo and Lin improved the overall quality of the subtitles by comparing and learning the correlation between the image and the subtitles. Through the model’s learning of “similar” data points and “negative samples,” the weight of the representation of similar data is gradually increased, thereby reducing the workload of sample labeling. He et al. used comparative learning to build a large dynamic dictionary and proposed comparative unsupervised learning to improve the detection ability of the model.

In view of the lack of information in the occlusion of underwater images, the existing image reconstruction technology cannot achieve image reconstruction. This article proposes fine reconstruction of underwater images with environmental feature fusion. In the construction of a fine reconstruction model, the attention mechanism will be embedded in the reconstruction model. At the same time, via positive sample and negative sample learning, the reconstruction model can pay attention to the salient features of the sample. The orthomorphic structure of the repair model in this article is shown in Figure 2. And the detailed principles will be elaborated in the third section.

Proposed approach

Feature extraction

Underwater objects tend to be obscured by foreign objects in complex underwater environments. In this article, the occluded underwater objects are defined as underwater targets. At the same time, the occluded area is defined as the missing area.

The underwater light is dim, and the movement of underwater plankton causes slight deformation of the underwater target. These factors undoubtedly increase the difficulty of underwater image feature extraction. At the same time, underwater images are difficult to obtain, which cannot meet the needs of deep learning models that require a large number of samples for training. To solve this problem, positive and negative sample learning is introduced to extract the features of underwater targets.

VGGNet, a network model proposed by Kamalyan and Simonyan, has the following characteristics of deep network, small convolution kernel, and small pooling kernel. Increasing the depth of the model can improve its feature extraction capability. At the same time, the model also has generalization capability to adapt to different data sets. Therefore, extract the features of underwater targets, using the visual geometry group (VGG)-16 model network. The class of the sample should be clarified. As shown in Figure 3, the underwater images with occlusion phenomenon are positive samples. The obscured underwater objects are defined as real values and the images irrelevant to the underwater target are defined as negative samples. The principle is shown in the following equation.
\[ S(f(x), f(x^+))^3 S(f(x), f(x^-)) \]  

(1)

where \( x^+ \) indicates samples similar to \( x \), called positive samples and \( x^- \) indicates samples not similar to \( x \), called negative samples. \( S(\cdot, \cdot) \) represents the degree of similarity between samples. Learning from positive and negative samples can make the difference between similar samples smaller and the difference between dissimilar samples larger. Then the positive sample is represented as \( \{x, x^+\} \). Negative samples represent \( \{x_1, x_2, ..., x_k\} \). To make the positive samples \( \{x, x^+\} \) have the concept of semantic level, this article classifies all samples with similar characteristics into e-type. Then the probability distribution of this class can be expressed as \( D_c \). Similarly, take \( D_b \) to represent the probability distribution of negative samples, which can be expressed as follows

\[
D_c(x, x^+) = E_{c \sim \rho} D_c(x)D_c(x^+) \tag{2}
\]

\[
D_b(x^-) = E_{b \sim \rho} D_b(x^-) \tag{3}
\]

where distribution \( D_c \) represents data similar to the target, distribution \( D_b \) represents data not related to the target, and \( \rho \) represents the distribution of data on the C category. Positive sample \( \{x, x^+\} \sim D_c \), negative sample \( \{x_1, ..., x_k\} \sim D_b \). According to the principle of Eq. (1), the model learned gradually approaches the positive sample and finally realizes the feature extraction of the underwater image.

The complex underwater environment results in underwater objects to be obscured. These factors that block objects are closely related to underwater objects. Such usage of the background information will contribute to the reconstruction of underwater images. This article proposes extracting relevant environmental information in the background by using hierarchical environmental feature attention mechanism. The environmental feature attention mechanism will retrieve and copy feature information patches from the known background to reconstruct the patch of the missing area. When selecting patches, the most concerned issue of this article is how to match the missing target features with the surrounding environment.

The paramount consideration is the local environmental characteristics, matching the missing pixel characteristics of the target with the surrounding environmental information. Extract the patch from the background and shape it into a convolution filter. To further verify the matching degree between the extracted background patch \( x' \) and the target features \( x \). This article uses cosine similarity to detect. The principle is shown in the following equation

\[
\hat{s}_c(x', x) = \frac{x \cdot x'}{\|x\| \cdot \|x'\|} \tag{4}
\]

where \( \hat{s}_c(x', x) \) represents the similarity between background patch \( x' \) and the target information \( x \). Then, the attention weight of each pixel in the softmax is weighed according to the similarity between the background patch \( x' \) and the target information \( x \). That is, the attention level of each pixel can be expressed as \( \hat{S}_c(x', x) = \alpha \text{softmax}_{x, x'}(\hat{s}_c(x', x)) \), where \( \alpha \) is a constant. The foreground feature changes equally with the change of the background patch corresponding to the attention. For example, the value of \( \hat{S}_c(x', x+1) \) is close. This article uses coherent thinking to achieve the consistency of attention to local environmental features. The principle is shown in the following equation

\[
\hat{s}_c(x', x) = \sum_{i \in \{k, ..., -k\}} S_c(x', x+1) \tag{5}
\]

To make better use of background information, this article proposes to aggregate global-level environmental feature information based on the overall characteristics of the input image. And the derived attention feature information is called global environmental attention feature information. The same as the above principle, the global-level environmental attention mechanism is expressed as follows

\[
\hat{S}_g(x', x) = \sum_{i \in \{k, ..., -k\}} S_g(x', x+i) \tag{6}
\]

Through the above, the background features of local environment attention and the background features of global environment attention can be obtained. In this article, the environment feature information of these two parts is fused by convolutional layer. Thereby a layered environmental attention mechanism is obtained. The degree of attention between the background patch \( x' \) and the target feature \( x \) can be expressed as follows

\[
s = f_{\text{conv}}\left(\hat{S}_c(x', x) \oplus \hat{S}_g(x', x)\right) \tag{7}
\]

where \( f_{\text{conv}}(\cdot) \) means convolution operation and \( \oplus \) means conjunction. The hierarchical environmental feature attention mechanism can make use of environmental information to a greater extent in the test, enriching the target data in the training.

**Refactoring model construction**

The model designed in this article is mainly divided into two parts: the rough reconstruction part and the fine reconstruction part. Its structure is shown in Figure 4. Input the occluded image \( I_{\text{in}} \) into the rough reconstruction network, and get a rough reconstructed image \( I_r \). The image \( I_{\text{in}} \) to be reconstructed and the rough reconstructed image \( I_r \) are put into the fine reconstructed network. The reconstructing network will quickly extract the effective feature information of the superimposed area. This can obtain a finely reconstructed image, so as to realize the reconstruction of the incomplete image.

The rough repair network designed in this article is a repair model based on GAN strategy. It associates each
layer of the encoder with the characteristics of the corresponding layer of the decoder. The encoder obtains the depth feature representation of the image to be reconstructed. And the decoder predicts and generates missing area information based on the feature. This article embeds the hierarchical environmental feature attention mechanism into the rough reconstruction network. In the image reconstruction model, Wasserstein Generative Adversarial Network Gradient Penalty (WGAN-GP) loss is better than the existing GAN loss, and it will produce a better effect when combined with the reconstruction loss function. The Wasserstein-1 distance in WGAN is expressed as follows when combined with the reconstruction loss function.

The WGAN-GP reconstruction model uses gradient penalty constraints to train and optimize the generator of the original WGAN-GP network. The generated data distribution is compared with the original data distribution by distance \( W(P_r, P_g) \). The following equation expresses its objective function

\[
\min_{G} \max_{D} \mathbb{E}_{\tilde{x} \sim P_r}[D(\tilde{x})] - \mathbb{E}_{z \sim P_g}[D(G(z))] \tag{10}
\]

where \( L \) represents the set of 1-Lipschitz functions. \( P_g \) is the implicit model distribution in \( \tilde{x} = G(z) \). \( G \) represents the input of the generator. The pixel-by-pixel reconstruction loss will directly return regresses holes to the current real image. WGAN learns to match the most likely correct image and applies the anti-gradient training generator. Since both losses measure distance in pixels, and combined losses are easier to train and make the optimization process more stable.

The image \( I_{in} \) and the rough reconstructed image \( I_r \) are put into the fine reconstruction network. The reconstructed network will quickly extract the effective feature information of the overlapped area. This greatly improves the reconfiguration effect of reconfigured networks. The network structure of the fine reconstruction model is similar to that of the rough reconstruction model. The difference is that the fine reconstruction network is embedded to...
enhance the feature correlation and feature continuity of the incomplete area. It can further improve the features of the reconstructed region by feature correlation. Make the incomplete part guessed more reasonable. The principle is shown in Figure 5.

Feature correlation is divided into two stages: search stage and generation stage. Generate each patch $x_{fi}$ for the reconstruction region $N (i \in (1,k), k$ denotes the number of small patches). The coherent layer of related features searches for the most matching patch $x_{fi}$ in the unoccluded area, which is used for initialization $x_{fi}$. Then uses patch $x_{fi}$ as the main feature information and refers to the already generated patch $x_{fi}$ to reconstruct three during the generation process. The principle can be expressed as follows

$$D_{max} = \frac{\langle x_{f1}, x_{f1} \rangle}{\|x_{f1}\| \cdot \|x_{f1}\|}$$ (11)

$$D_a = \frac{\langle x_{f1}, x_{f1-1} \rangle}{\|x_{f1}\| \cdot \|x_{f1-1}\|}$$ (12)

where $D_a$ represents the similarity between this neighboring patch. $D_{max}$ represents the similarity between the best matching patch $x_{f1}$ and the context area patch $x_{f1}$. Consider $D_a$ and $D_{max}$ as the weight of the generated patch. So that each patch contains the relevant characteristics of the feature and the information of the previous patch. Therefore, the generated patch can be expressed as follows

$$x'_{f1} = \frac{D_{at}}{D_{at} + D_{maxi}} \times x_{f1-1} + \frac{D_{maxi}}{D_{at} + D_{maxi}} \times x_{f1}, i \in (2,k)$$

(13)

The patch generation process is an iterative process, so $x'_{f1}$ is related to all previous patches $x'_{f1-1}$ and $x'_{f1}$. Each patch obtained has simultaneously obtained more information on feature correlation. Finally, the patch extracted in $X_f$ is used as a deconvolution filter to reconstruct $X_f$ to obtain a more realistic repair result. To further improve the effect of image reconstruction, a feature reconstruction recognizer is introduced in this article. The recognizer can distinguish between the original image and the reconstructed image. The adversarial loss is calculated based on the information of the special reconstruction features. The adversarial loss function $D_R$ is used for fine repairing the network, and the loss function $D_M$ is used for rough repairing the network as shown in the following equation

$$\begin{cases}
D_R = -E_{in\epsilon}(1 - D(I_{in}, I_r))^2 - E_{in\epsilon}[D(I_r, I_{in})^2] \\
D_M = -E_{in\epsilon}[D(I_m, I_{in})^2] - E_{in\epsilon}[D(I_{in}, I_m)^2]
\end{cases}$$ (14)

where $D$ represents the discriminator. $E_{in\epsilon}$ represents the operation of averaging the original data after discriminator identification. $E_{in\epsilon}$ represents the operation of averaging the generated data after discriminator identification.

**Image reconstruction and loss function**

The extraction of underwater image features is difficult. Also underwater images are difficult to obtain, which cannot meet the requirements of deep learning models that require a large number of samples for training. The VGG-16 model pretrained in the stereo quantitative underwater image data set (SQUID) database. The reconstruction model extracts feature information of the image to be reconstructed by learning positive and negative samples.
And the feature representation of the sample data is learned using the data itself as supervised information. Then the objective function is shown in the following equation

$$L_d = E \left[ \ell \left( \left\{ f(x)^T f(x^c)^T - f(x^+) \right\}_{i=1}^k \right) \right]$$

Thus, the relationship between \(M\) samples from \(D_z \times D_n\) can be expressed as follows

$$L'_d = \frac{1}{M} \sum_{i=1}^{M} \ell \left( \left\{ f(x)^T f(x^c)^T - f(x^+) \right\}_{i=1}^k \right)$$

Through the above analysis, the objective function based on unsupervised learning can be expressed as follows

$$L_z = E_{c,c'} E_{y,y'} \left[ \ell \left( \left\{ f(x)^T f(x^c)^T - f(x^+) \right\}_{i=1}^k \right) \right]$$

Image reconstruction often used perceptual loss to improve the reconstruction ability of image reconstruction networks. Since the reconstructed model in this article contains the relevant feature coherent layer. The past perceptual loss methods cannot be directly used to optimize the repair model based on convolution. Otherwise, it will affect the training of the reconstructed model and the reconstructed results. To solve this problem, this article adjusts the form of perceptual loss and proposes a solution to consistency loss. Set the encoder and the decoder feature spaces corresponding to the area to be reconstructed as the target. And the distance \(L_c\) is calculated according to this theory. The resulting consistency loss is shown in the following equation

$$L_c = \sum_{i \in M} \left\| W(I_w)_z - \phi_m(I_o)_z \right\|_2 + \left\| W_d(I_w)_z - \phi_m(I_o)_z \right\|_2$$

where \(\phi_m\) represents the parameters obtained through training in the VGG-16 network model. \(W_d(\cdot)\) represents the feature of the coherent layer of related features in the encoder, and \(W_m(\cdot)\) is the corresponding layer phase feature space of the relevant feature coherent layer in the decoder.

To make the rough repair image \(I_r\) and the fine reconstructed image \(I_m\) closer to the real image, we will use the distance \(L_1\) as our image reconstruction loss. The loss function is shown in the following equation

$$L_r = \|I_r - I_o\|_1 + \|I_m - I_o\|_1$$

Taking into account the consistency loss, edge loss function, contrastive loss, and reconstruction loss, the overall goal of the reconstruction model in this article is defined as follows

$$L = \alpha_r L_r + \alpha_c L_c + \alpha_d L_d + \alpha_h L_h$$

where \(\alpha_r\), \(\alpha_c\), \(\alpha_d\), and \(\alpha_h\) are the loss parameters of consistency loss, edge loss function, contrastive loss, and reconstruction loss, respectively.

### Experiment

In this section, a large number of experiments are designed to verify the performance of our proposed image reconstruction model. And comparative experiment is conducted between the current article and the existing methods in SQUID,36 Real-world Underwater Image Enhancement (RIUE) data set,37 and underwater image data set.

### Experimental setup

The experiment designed in this article is based on three data sets: SQUI data set, RIUE data set, and underwater image data set. The learning rate is set to \(2 \times 10^{-4}\) and \(\beta = 0.05\). The trade-off parameters are set to \(\alpha_r = 0.5\), \(\alpha_d = 0.1\), \(\alpha_c = 0.1\), and \(\alpha_h = 0.001\). We trained the raw data of the three data sets for 3 days, 3 days, and 1 day, respectively. All used for training and testing are 256 × 256, and the running speed on the GPU is 0.82 s/frame. We will compare the four methods, namely GAN,21 CE,25 CA,27 and coherent semantic attention (CSA).28 The hardware configuration of the experiment is as follows: CPU is Intel(R) Core(TM) i7-8700K@3.70 GHz and GPU is RTX 2080 Ti. Memory is 64G. The running environment is Python3.7, the model is written using the PyTorch library, and the operating system is Ubuntu-16.04.

### Qualitative experimental comparison

SQUID is a data set mainly aiming at the reconstruction of various underwater targets and underwater scenes. It is widely used in the field of underwater image research. It is mainly used in research fields such as underwater image enhancement, underwater image defogging, and underwater image three-dimensional (3D) reconstruction. This article selects a large number of underwater figures from the SQUID. The input image is occluded by the splash of the target feature, and the occluded part is reconstructed by the algorithm proposed in this article. In the experiment, the proposed algorithm is compared with GAN, CE, CA, and CSA image reconstruction models, with the simulation result shown in Figure 6.

The simulation results are shown in Figure 6. The input in the leftmost column is the underwater image where the target feature is lost due to splashes or water ripples. Others are the reconstruction results of GAN, CE, CA, and CSA algorithms. Experiment (A), experiment (B), experiment (D), and experiment (E) are experiments in which target features are lost due to water spray. Experiment (C) is an experiment of target deformation caused by water ripples. It can be obtained from the experimental results that GAN, CE, CA, and CSA cannot effectively reconstruct underwater images. For example, in experiment (A)
and experiment (D), although the CA and CSA models can be important the occluded part of the structure target, there is a fuzzy texture. The algorithm proposed in this article outperforms other algorithms in most cases. The model reconstruction results in this article are relatively clear and can effectively reconstruct the occlusion information. For example, experiment (C) and experiment (E) can reconstruct the information of human occlusion area well. And there is no loss of target features. Since the underwater target obscures too much area, there may be not enough learning of the scene to learn more features about the human body. Therefore, the image reconstruction is not so satisfactory. However, the restoration of the proposed algorithm is superior to other algorithms to some extent.

**Quantitative experimental analysis**

Due to individual differences, personal preferences, and other subjective factors, the evaluation of experimental results will be one-sided to a certain extent. To obtain a more accurate quality evaluation of the repair results, this article introduces peak signal-to-noise ratio (PSNR)\(^{38}\) and structural similarity index (SSIM)\(^{39}\) to analyze the repair results from the data to get more accurate evaluation results. The signal-to-noise ratio evaluates the quality of the image through the error of the pixels corresponding to the two images. The larger the value, the better the repair result of the image

\[
MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \|x(i,j) - y(i,j)\|^2
\]  

(21)

\[
PSNR = 10 \cdot \log_{10} \left( \frac{MAX^2}{MSE} \right)
\]  

(22)

where \(m\) and \(n\) represent the size of the image, and MSE represents the mean square error between the images. MAX\(^2\) represents the maximum value taken in the image.

Structural similarity is important to measure the structural similarity of the image from the brightness, contrast, and structural information of the image, thereby evaluating the degree of distortion of the image. The larger the value, the smaller the distortion. The closer the repaired image is...
Ours & 21.092 & 21.044 & 19.574 & 26.827 & 20.125 \\
CSA & 20.984 & 20.041 & 19.139 & 26.801 & 20.047 \\
CA & 20.163 & 19.986 & 18.299 & 26.784 & 19.938 \\
CE & 19.998 & 19.463 & 17.519 & 26.715 & 19.663 \\
GAN & 19.690 & 19.039 & 17.512 & 26.699 & 19.609 \\

### Table 2. SSIM values under different algorithms of SQUID database.

| Methods | Experiment number |
|---------|-------------------|
|        | A    | B    | C     | D     | E     |
| GAN    | 0.965 | 0.965 | 0.882 | 0.926 | 0.946 |
| CE     | 0.971 | 0.971 | 0.879 | 0.912 | 0.953 |
| CA     | 0.973 | 0.975 | 0.889 | 0.945 | 0.971 |
| CSA    | 0.975 | 0.980 | 0.927 | 0.951 | 0.972 |
| Ours   | 0.986 | 0.985 | 0.986 | 0.973 | 0.978 |

SQUID: stereo quantitative underwater image data set; SSIM: structural similarity index; GAN: generative adversarial network; CE: contextual encoder; CA: contextual attention; CSA: coherent semantic attention.

### Methods

| Experiment number |
|-------------------|
| A    | B    | C     | D     | E     |

#### Table 1. PSNR values under different algorithms of SQUID database.

| Methods | Experiment number |
|---------|-------------------|
|        | A    | B    | C     | D     | E     |
| GAN    | 19.690 | 19.039 | 17.512 | 26.699 | 19.609 |
| CE     | 19.998 | 19.463 | 17.519 | 26.715 | 19.663 |
| CA     | 20.163 | 19.986 | 18.299 | 26.784 | 19.938 |
| CSA    | 20.984 | 20.041 | 19.139 | 26.801 | 20.047 |
| Ours   | 21.092 | 21.044 | 19.574 | 26.827 | 20.125 |

PSNR: peak signal-to-noise ratio; SQUID: stereo quantitative underwater image data set; GAN: generative adversarial network; CE: contextual encoder; CA: contextual attention; CSA: coherent semantic attention.

#### to the original image, the principle is shown in the following equation

$$SSIM = \frac{(2\sigma_x \sigma_y + C_1)(2\delta_{xy} + C_2)}{(\sigma_x^2 + \sigma_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where $\sigma_x$ and $\sigma_y$ are the average values of the original image and the restored image, respectively. $\sigma_x^2$ and $\sigma_y^2$ represent the variance. $\delta_{xy}$ represents the covariance. $C_1$ and $C_2$ represent the constant. Then the PSNR and SSIM of the above experiment can be expressed in the following.

When evaluating the quality of image reconstruction, people often use PSNR and SSIM for evaluation. The PSNR value of Figure 6 is shown in Table 1. It can be seen from the table that the repair result of this model is better than other models. The PSNR value of experiment (A) and experiment (B) is significantly higher than that of experiment (C), experiment (D), and experiment (E). The SSIM value of Figure 6 is shown in Table 2. It can be seen from the table that the SSIM value of the repair result of the model in this article is higher than that of other models. The PSNR value is significantly higher than that of other groups of experiments, thus explaining the underwater scene. It has a great influence on underwater image reconstruction.

RUIE data set is a data set for underwater image research proposed by Dhirubhai Ambani Institute of Information and Communication Soft International Institute. RUIE data set has several features: rich underwater images, large data volume of its main feature image, image scenes, various colors, rich detection targets, etc. It is mainly used for detection and recognition of underwater targets, and enhancement and recovery of underwater images. This article selects a large number of frogman images from the RUIE database. The local features of these frogmen are obscured by fish and then the target reconstruction for the obscured part is achieved by the proposed algorithm. In the experiments, the algorithm of this article is compared with GAN, CE, CA, and CSA image reconstruction models. The simulation results are shown in Figure 7 and their PSNR values and SSIM values are presented in Tables 3 and 4, respectively.

The simulation result is shown in Figure 6. The input in the leftmost column is the underwater image where the target feature is lost due to the occlusion of fish. Others are the reconstruction results of GAN, CE, CA, CSA, and ours image reconstruction models in order. It can be obtained from the experimental results that GAN, CE, CA, and CSA cannot effectively reconstruct underwater images. For example, in experiment (A), experiment (D), and experiment (E), the results of GAN and CE model reconstruction lose a lot of target features. As in experiments (D) and (E), although the CA and CSA models are able to reconstruct the obscured part of the target, there are blurred textures and the reconstruction area is blurred. The results of the model reconstruction in this article are relatively clear and can effectively reconstruct the occlusion information. As in experiment (D) and experiment (E), the algorithm proposed in this article can over effectively reconstruct the features of the human body, while other reconstructive models cannot. The experimental results verify that the image reconstruction model proposed in this article has certain advantages in the underwater field.

The PSNR values of the experimental results are presented in Table 5. From the data, it can be seen that the reconstruction results of this model are better than other models, and the PSNR values of experiments (E) and (F) are significantly higher than those of other experimental groups. It can be seen from the table that the SSIM value of the reconstruction result of the model in this article is significantly higher than that of other models, and the SSIM value is significantly higher than that of other groups of experiments. This shows that the proposed reconstruction model has certain advantages in underwater image reconstruction. To further demonstrate the superiority of the proposed algorithm, the PSNR and SSIM values corresponding to the experimental results of Figure 7 are represented by the line graphs as shown in Figures 8 and 9, respectively.
Figure 7. Repair results of each model in the RUIE database. The experiment is aimed at a scenario where the frogmen are obscured by fish, and the experiment aims to reconstruct the features of the frogmen in the obscured area. The leftmost row of the figure is the input image, and the others are the reconstruction results of GAN, CE, CA, CSA, and our image reconstruction models in order. RUIE: real-world underwater image enhancement; GAN: generative adversarial network; CE: context encoder; CA: contextual attention; CSA: coherent semantic attention.

Table 3. PSNR value of each repair result of RUIE database.

| Methods | Experiment number |
|---------|------------------|
|         | A    | B    | C    | D    | E    | F    |
| GAN     | 16.081| 17.779| 18.491| 16.159| 18.145| 21.617|
| CE      | 17.532| 18.176| 18.523| 16.384| 18.154| 21.818|
| CA      | 18.114| 18.845| 19.401| 17.489| 18.863| 22.325|
| CSA     | 18.667| 19.245| 19.603| 17.493| 19.066| 22.506|
| Ours    | 18.989| 20.653| 19.869| 17.885| 19.192| 22.989|

PSNR: peak signal-to-noise ratio; RUIE: real-world underwater image enhancement; GAN: generative adversarial network; CE: context encoder; CA: contextual attention; CSA: coherent semantic attention.

Table 4. PSNR value of each repair result of RUIE database.

| Methods | Experiment number |
|---------|------------------|
|         | A    | B    | C    | D    | E    | F    |
| GAN     | 0.941 | 0.939 | 0.954 | 0.919 | 0.874 | 0.935 |
| CE      | 0.947 | 0.951 | 0.961 | 0.924 | 0.881 | 0.939 |
| CA      | 0.952 | 0.965 | 0.968 | 0.928 | 0.885 | 0.943 |
| CSA     | 0.963 | 0.971 | 0.972 | 0.936 | 0.889 | 0.948 |
| Ours    | 0.971 | 0.979 | 0.985 | 0.942 | 0.913 | 0.966 |

PSNR: peak signal-to-noise ratio; RUIE: real-world underwater image enhancement; GAN: generative adversarial network; CE: context encoder; CA: contextual attention; CSA: coherent semantic attention.
Algorithm verification

This article will further verify the repair model proposed in this article on the underwater target data set. The underwater target data set is a data set established by our team’s laboratory mainly for underwater images. It is of great significance in the research of underwater image processing. These include torpedoes, submarines, frogmen, AUVs, and other categories. It is currently under further construction and improvement. This article selects a large number of occluded or blurred torpedo, submarine, and AUV images from the underwater target data set for experiments and compares them with GAN, CE, CA, and CSA image reconstruction models. The simulation results are shown in Figure 10.

The simulation result is shown in Figure 10. The input in the leftmost column is the underwater image where the target feature is lost due to the occlusion of fish. The results of GAN and CE image reconstruction model reconstruction have the phenomenon of target loss. For example, the reconstruction results of experiment (A) and experiment (C) are not ideal. In experiment (C), although the reconstruction results of CA and CSA can reconstruct part of the contour, the result is not very satisfactory. The reconstruction algorithm proposed in this article can effectively reconstruct underwater targets.

The PSNR value of Figure 10 is shown in Table 5. It can be seen from the table that the reconstruction results of this model are better than other models. The PSNR values of experiment (D) and experiment (E) are significantly higher than those of experiment (A), experiment (B), and experiment (C). The SSIM values of Figure 10 are presented in Table 6. It can be seen from the table that the SSIM and PSNR values of the reconstruction results of this model are higher than those of other models, which show that underwater scenes have a great influence on underwater image reconstruction. Experimental data are displayed more intuitively with 3D histograms in this article. As shown in Figures 11 and 12, it can be seen from the simulation results that the reconstruction results of the proposed model are better than other models for comparison. No matter it is a qualitative comparison or a quantitative comparison, the proposed algorithm is superior than other comparative models, as for the underwater image reconstruction with missing target information.

### Table 5. PSNR value of each repair result in the underwater target data set.

| Methods | Experiment number |
|---------|-------------------|
| GAN     | 18.927 18.803 15.448 16.989 21.047 |
| CE      | 19.199 18.775 15.494 16.701 21.242 |
| CA      | 19.328 18.881 16.198 18.326 21.640 |
| CSA     | 20.034 18.994 17.131 18.331 22.341 |
| Ours    | 20.141 19.348 17.499 18.647 22.903 |

PSNR: peak signal-to-noise ratio; GAN: generative adversarial network; CE: context encoder; CA: contextual attention; CSA: coherent semantic attention.

### Table 6. SSIM values of the repair results in the underwater target data set.

| Methods | Experiment number |
|---------|-------------------|
| GAN     | 0.902 0.933 0.921 0.951 0.947 |
| CE      | 0.901 0.945 0.924 0.954 0.951 |
| CA      | 0.914 0.951 0.928 0.964 0.959 |
| CSA     | 0.917 0.957 0.931 0.967 0.963 |
| Ours    | 0.925 0.964 0.947 0.973 0.982 |

SSIM: structural similarity index; GAN: generative adversarial network; CE: context encoder; CA: contextual attention; CSA: coherent semantic attention.
Conclusion

This article proposes an underwater image reconstruction algorithm based on environment feature fusion. Firstly, the significant features of the image are extracted by positive and negative sample learning. Secondly, the relevant information in the background is retrieved by the environmental attention mechanism. Finally, a coarse-to-fine underwater image restoration model is constructed to obtain fine

Figure 10. Repair results of each model in the underwater target data set. The experiment is aimed at the occlusion and blur phenomenon of underwater AUV and other targets and reconstructs the occlusion and blur area. The leftmost row of the figure is the input image, and the others are the reconstruction results of GAN, CE, CA, CSA, and ours image reconstruction models in order. AUV: autonomous underwater vehicle; GAN: generative adversarial network; CE: context encoder; CA: contextual attention; CSA: coherent semantic attention.

Figure 11. PSNR value of each repair result in the underwater target data set. PSNR: peak signal-to-noise ratio.

Figure 12. SSIM value of each repair result in the underwater target data set. SSIM: structural similarity index.
restoration results. The test results show that the proposed algorithm has certain advantages in both qualitative and quantitative comparisons. However, proposed the algorithm has obvious shortcomings for the heavily obscured underwater objects. The performance of the algorithm will be further improved in future research work.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article is supported by the National Key Research and Development Project [2019YFB1311000], the Science and Technology Project of Henan Province [202102210302] and [182102210302], the Key Scientific Research Projects of the Institutions of Higher Education in Henan Province [21A520013], and the National Defense Science and Technology Innovation Special Zone.

ORCID iDs
Jiyong Zhou https://orcid.org/0000-0003-3124-8394
Lei Cai https://orcid.org/0000-0003-4811-5854

References
1. Tengyue L, Shenghui R, Xueting C, et al. Underwater image enhancement framework and its application on an autonomous underwater vehicle platform. Opt Eng 2020; 59(8): 1–10.
2. Wynn RB, Huvenne VAI, Murton B, et al. Autonomous underwater vehicles (AUVs): their past, present and future contributions to the advancement of marine geoscience. Mar Geol 2014; 352: 451–468.
3. Pfeffer A, Wu C, Fry G, et al. Software adaptation for an unmanned undersea vehicle. IEEE Softw 2019; 36(2): 91–96.
4. Raihan JA, Abas PE, De Silva LC, et al. Review of underwater image restoration algorithms. IET Image Process 2019; 13(10): 1587–1596.
5. Garcia R, Nicosevici T, Gracias N, et al. Exploring the seafloor with underwater robots: land, sea & air. In: Computer vision in vehicle technology. Chapter 4, in Land, Sea & Air (eds AM Lopez, A Imaiya, T Pajdla, and AM Alvarez), 10 February 2017, pp. 75–99.
6. Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting. In: Proceedings of the 27th annual conference on computer graphics and interactive techniques, New York, NY, USA, 1 July 2000, pp. 417–424.
7. Zhaoyi Y, Xiaoming L, Mu L, et al. Shift-net: image inpainting via deep feature rearrangement. In: Proceedings of the European conference on computer vision, Munich, Germany, 08–14 September 2018, pp. 3–19.
8. Xiong W, Jiahui Y, Lin Z, et al. Foreground-aware image inpainting. In: IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 16 June–20 June 2019, pp. 5833–5841.
9. Nazeri K, Ng E, Joseph T, et al. EdgeConnect: generative image inpainting with adversarial edge learning. In: IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 16–20 June 2019, pp. 5833–5841.
10. Efros AA and Leung TK. Texture synthesis by nonparametric sampling. In: IEEE international conference on computer vision, Seoul, Korea, 27 October–2 November 2019, pp. 5833–5841.
11. Marcello B, Luminita V, Guillermo S, et al. Simultaneous structure and texture image inpainting. IEEE Trans Image Process 2003; 13(9): 882–889.
12. Criminisi A, Perez P, Toyama K, et al. Region filling and object removal by exemplar-based image inpainting. IEEE Trans Image Process 2004; 13(9): 1200–1212.
13. Le MO, Gautier J, and Guillemot C. Exemplar-based inpainting based on local geometry. In: IEEE international conference on image processing, Brussels, Belgium, 11–14 September 2011, pp. 3401–3404.
14. David GL. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004; 60(2): 91–110.
15. Chan TF and Shen J. Non-texture inpainting by curvature-driven diffusions. J Vis Commun Image Represent 2001, 12(4): 436–449.
16. Chan TF, Shen J, and Vese L. Variational PDE models in image processing. Not Am Math Soc 2002; 50(1): 14–26.
17. Huang JB, Kang SB, Ahuja N, et al. Image completion using planar structure guidance. Trans Graph 2014; 33(4): 1–10.
18. Xie J, Xu L, and Chen E. Image denoising and inpainting with deep neural networks. In: International conference on neural information processing systems, Seoul, Korea, 27 October–2 November 2019, pp. 341–349.
19. Satoshi I, Simo-Serra E, and Ishikawa H. Globally and locally consistent image completion. Trans Graph 2017; 36(4): 1–14.
20. Dekel T, Gan C, Krishnan D, et al. Sparse, smart contours to represent and edit images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018, pp. 3511–3520.
21. Goodfellow J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in neural information processing systems, Montreal, Canada, 8–13 December 2014, pp. 2672–2680.
22. Zheng C, Cham TJ, and Cai J. Pluralistic image completion. In: IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 16–20 June 2020, pp. 1438–1447.
23. Zeng Y, Fu J, Chao H, et al. Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–21 June 2015, pp. 2048–2057.
24. Barnes CE, Shechtman A, Goldman DB, et al. Patchmatch: a randomized correspondence algorithm for structural image editing. Trans Graph 2009; 28(3): 1–11.
25. Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: feature learning by inpainting. In: IEEE conference on
26. Chao Y, Xin L, Zhe L, et al. High-resolution image inpainting using multi-scale neural patch synthesis. In: *IEEE conference on computer vision and pattern recognition*, Honolulu, Hawaii, USA, 21 July–26 July 2017, pp. 4076–4084.

27. Jiahui Y, Zhe L, Jimei YX, et al. Generative image inpainting with contextual attention. In: *IEEE conference on computer vision and pattern recognition*, Salt Lake City, Utah, USA, 18–22 June 2018, pp. 5505–5514.

28. Liu H, Jiang B, Xiao Y, et al. Coherent semantic attention for image inpainting. In: *IEEE/CVF international conference on computer vision*, Seattle, WA, USA, 16–20 June 2020, pp. 4169–4178.

29. Zhaomao C, Xin J, Borui Z, et al. Hierarchical context embedding for region-based object detection. In: *European conference on computer vision*, Glasgow, UK, 23–28 August 2020, pp. 633–648.

30. Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention. In: *Advances in neural information processing systems*, Montreal, Canada, 8–13 December 2014, pp. 2204–2212.

31. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: *Computer science*, Lille, France, 6–11 July 2015, pp. 2048–2057.

32. Bo D and Lin D. Contrastive learning for image captioning. In: *Computer vision and pattern recognition*, Honolulu, Hawaii, USA, 21–26 July 2017, pp. 898–907.

33. Sanjeev A, Hrishikesh K, Mikhail K, et al. A theoretical analysis of contrastive unsupervised representation learning. In: *International conference on machine learning*, Long Beach, USA, 9–15 June 2019, pp. 9904–9923.

34. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In: *IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 16–20 June 2020, pp. 9726–9735.

35. Kamalyan AG and Simonyan VV. On the number of solutions of a certain type of one-dimensional pseudodifferential equations in the Sobolev-Slobedetski space. *J Shellfish Res* 2002, 21(1): 201–210.

36. Li C, Guo C, Ren W, et al. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans Image Process* 2020; 29: 4376–4389.

37. Liu RS, Fan X, Zhu M, et al. Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light. *IEEE Trans Circuits Syst Video Technol* 2020; 30(12): 4861–4875.

38. Tuanji W, Jianhua L, Yi L, et al. Image quality evaluation based on image weighted separating block peak signal to noise ratio. In: *International conference on neural networks & signal processing*, Nanjing, China, 14–17 December 2003, pp. 994–997.

39. Tuncel E, Ferhatosmanoglu H, and Rose K. VQ-index: an index structure for similarity searching in multimedia databases. In: *Electrical and computer engineering University of California*, Juan-les-Pins, France, 1–6 December 2002, pp. 543–552.