This is a repository copy of *Three-dimentional reconstruction of underwater side-scan sonar images based on shape-from-shading and monocular depth fusion*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/209353/

Version: Published Version

imts Intelligent Marine
Technology and Systems

Check for
updates

Three-dimentional reconstruction of underwater side-scan sonar images based on shape-from-shading and monocular depth fusion

Yakun Ju[1*] , Jingchun Zhou[2], Shitong Zhou[3], Hao Xie[4], Cong Zhang[4], Jun Xiao[4], Cuixin Yang[4] and Jianyuan Sun[5]

Abstract

Modern marine research requires high-precision three-dimensional (3D) underwater data. Underwater environments experience severe visible light attenuation, which leads to inferior imaging compared with air. In contrast, sound waves are less affected underwater; hence side-scan sonar is used for underwater 3D reconstruction. Typically, the shape-from-shading algorithm (SfS) is widely used to reconstruct surface normal or heights from side-scan sonar images. However, this approach has challenges because of global information loss and noise. To address these issues, this study introduces a surface-normal fusion method. Specifically, we propose a frequency separation SfS algorithm using a discrete cosine transform, which provides a surface-normal map with less noise. We then fuse the surface-normal map with a novel depth estimation network to achieve high-precision 3D reconstruction of underwater side-scan sonar images. We conducted experiments on synthetic, NYU-depth-v2, and real side-scan sonar datasets to demonstrate the effectiveness of the proposed method.

Keywords Side-scan sonar, Shape-from-shading, Underwater 3D reconstruction, Depth estimation

1 Introduction

Recently, ocean exploration and utilization of oceanic resources and energy have gained significance worldwide (Zhou et al. 2023a, b). Modern marine research has an urgent requirement for high-precision three-dimensional (3D) underwater data (Fan et al. 2017). Underwater 3D reconstruction primarily relies on two key technologies: optical and acoustic imaging. Underwater visibility is substantially limited due to the attenuation of visible light, decreasing the effectiveness of optical imaging underwater compared with air.

In contrast, acoustic imaging has found extensive applications in underwater measurements, seabed operations, underwater archaeology, underwater navigation, and various other fields because of its minimal underwater attenuation and long-distance propagation capabilities. Side-scan sonar technology has been widely used for seabed terrain and object reconstruction. This method operates by emitting sound pulse waves to the left and right sides through sound wave transducer arrays located on either side of the device (Key 2000). When these emitted sound waves encounter water or objects on the seabed, they scatter in both directions, generating echoes. These echoes are captured by the

*Correspondence:
Yakun Ju
kelvin.yakun.ju@gmail.com
[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
[2] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
[3] Agricultural Bank of China Shaanxi Branch, Xi'an 710000, China
[4] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
[5] Department of Computer Science, University of Sheffield, Sheffield S14DP, UK
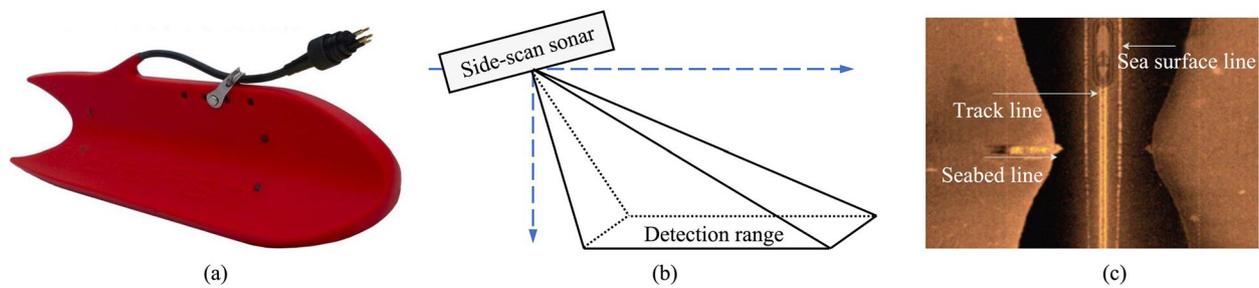
 Springer

Fig. 1 **a** Side-scan sonar 'StarFish' from BlueView company. **b** Schematic of the side-scan sonar scanning range. **c** Example of the captured sonar image

sonar and subsequently converted into electrical signals through a series of processes. The intensity of the echoes is displayed on the sonogram corresponding to variations in brightness within the pixels of the side-scan sonar image. Stronger echoes are represented by higher gray values in the sonar image. Figure 1 illustrates the 'StarFish' side-scan sonar and the captured sonar image.

The imaging model of sonar is similar to optical reflection, and researchers often simplify the sonar imaging model by approximating it to the Lambertian model (Durá et al. 2004; Coiras et al. 2007). The Lambertian model assumes that the object's surface exhibits diffuse reflection characteristics, where the reflection intensity depends solely on the angle between the incident light and the surface's normal direction, independent of the reflection direction (Ju et al. 2021, 2023c). Shape-from-shading (SfS) algorithm (Horn 1970) can be applied to deduce the object's 3D shape using the Lambertian reflection model. In other words, by knowing the gray-scale variations within a single image, these methods calculate the object's surface normal using the Lambertian model and subsequently derive its surface height. However, this approach has challenges because of the global information loss and noise.

To address the above challenges, we explored the following aspects.

(1) High-low frequency separation SfS: Traditional SfS algorithms (Horn 1970; Ikeuchi and Horn 1981; Frankot and Chellappa 1988) tend to overly emphasize reconstructing high-frequency details while overlooking the low-frequency information of the overall scene. To overcome these limitations, we employ the discrete cosine transform (DCT) (Ahmed et al. 1974) to split the sonar images into high- and low-frequency components. The low-frequency image is then reconstructed using the SfS algorithm, whereas the high-frequency image contains too much noise and is discarded. Our proposed high-low frequency separation SfS method enhances the reconstruction accuracy of traditional SfS methods.

(2) Depth estimation network with dilated convolution and attention mechanism: We also propose a monocular depth estimation network that combines the attention mechanism (Vaswani et al. 2017) with dilated convolution (Yu et al. 2017). The proposed network introduces a multiscale attention mechanism within the connection between the encoder and decoder. This enables the extraction of detailed features in the scene, ultimately improving depth estimation.

(3) Normal-depth fusion algorithm: We introduce a normal-depth fusion algorithm to fuse the surface normal maps obtained from (1) and the depth maps obtained from (2). The depth map produced using the depth estimation algorithm places a stronger emphasis on global depth changes, which contrasts with the focus of SfS. By combining these two methods, we address each other's limitations, resulting in a 3D reconstruction that effectively balances global and local details.

We conducted experiments on synthetic (Gwon et al. 2017), NYU-depth-v2 (Silberman et al. 2012), and real side-scan sonar datasets to demonstrate the effectiveness of the proposed normal-depth fusion method. The remainder of this paper is organized as follows: Section 2 discusses related works. Section 3 presents the proposed method. The experimental results are presented in Section 4.

2 Related work

2.1 Shape-from-Shading (SfS)

SfS proposed by Horn (1970) is a classic algorithm used for monocular image-based 3D reconstruction. The SfS algorithm operates by deducing the 3D shape of the target, relying on the assumption of the Lambertian model. In contrast to photometric stereo methods like Ju et al. (2020a, 2022, 2023a, b); Liu et al. (2022), SfS uses a single image as input, making it capable of addressing surface reconstruction in dynamic scenes with non-rigid

Ju *et al. Intelligent Marine Technology and Systems*   (2024) 2:4

Page 3 of 15

objects. SfS can be categorized into three types based on the reconstruction process: cost function minimization, propagation, and local methods.

Cost function minimization methods are commonly used to address the Lambert reflection reconstruction problem by iteratively adjusting the cost function (Ikeuchi and Horn 1981). This problem involves determining the surface normal direction in the *x*- and *y*-axes, given the grayscale of the input image. Various constraints have been introduced to effectively solve this problem. For example, Brooks and Horn (1985) introduced two fundamental constraints: photometric consistency and smoothness. Photometric consistency ensures that the grayscale of the inverted 3D object reconstruction matches that of the input image. Smoothness constraints maintain a gradual and natural transition along the edges of the reconstructed surface. Frankot and Chellappa (1988) introduced integrability constraints to ensure that the reconstructed surface is integrable and accurately represents the real scene. For specific tasks, such as side-scan sonar imaging, Zheng and Chellapa (1991) proposed a multiresolution 3D reconstruction method that uses a coarse-to-fine structure.

Propagation methods involve calculating other points along characteristic splines based on the depth and normal direction of a known point. Horn (1970) started from a singular point and propagated along the surface curve of the surrounding spherical neighborhood. The propagation direction follows the change in surface gradient. Rouy and Tourin (1992) introduced an SfS method based on the Hamilton-Jacobi-Bellman equation (Bellman 1966) and the viscous solution theory. They applied dynamic programming to establish the relationship between the viscous solution and optimal control, enabling the retrieval of a unique solution for SfS. Kimmel and Bruckstein (1995) employed multilayer contours initially defined by closed curves near singular points to reconstruct the object's surface.

Local methods recover the shape information of the target by analyzing the image's brightness and its first and second derivatives. Pentland (1984) assumes that the surface where each point in the image resides can be approximated as a local sphere. Lee and Rosenfeld (1985) leveraged this local spherical assumption to determine the inclination and declination of the object surface's normal with respect to the base light source coordinate system by analyzing the first derivative of the image's brightness.

2.2 Monocular depth estimation
Deep learning techniques have shown great capability in many computer vision tasks (Ju et al. 2020b; Kong et al. 2022; Rao et al. 2023; Xiao et al. 2023; Yu et al. 2023; Zhang et al. 2023b). Monocular depth estimation,

driven by deep neural networks (Wang et al. 2020; Kong et al. 2021; Luo et al. 2021; Xiao et al. 2021; Zhang et al. 2023a), has recently seen significant advancements because of its potent feature learning capabilities. Monocular depth learns the depth map of an object from a single input image. Eigen et al. (2014) first used a coarse-to-fine network for monocular depth estimation. In this approach, the former part of the network learns the global depth of the scene, while the latter optimizes the local details of the target shape. Li et al. (2015) introduced a dual-process framework based on the VGG-16 model. These two processes estimate the depth and gradient values of the object, respectively. Chen et al. (2016) developed a multiscale network to predict the depth value of each pixel by learning the relative depth of the scene. The network employs the relative depth error as the loss function. Recently, Lee et al. (2019) proposed a multiscale planar guidance layer as a replacement. This layer adopts local planar assumptions to directly obtain the original resolution feature map. Gan et al. (2018) employed a transformation layer to reflect the positional relationships between images and estimate the relative depth information between different objects in the scene through the position changes in adjacent images. Jung et al. (2017) employed generative adversarial networks (GANs) for single-image depth value prediction. The generator combines GlobalNet to extract global features and RefinementNet to estimate local details.

3 Proposed method
In this section, we present the proposed normal-depth fusion method for underwater side-scan sonar image 3D reconstruction. Figure 2 shows the overview framework. The proposed method can be divided into three parts: high-low frequency separation SfS method (Section 3.1), local planar guidance monocular depth estimation network (Section 3.2), and normal-depth fusion algorithm (Section 3.3).

3.1 High-low frequency separation SfS method
First, we propose an improved SfS surface reconstruction method for underwater side-scan sonar images. Traditional SfS techniques have demonstrated good capability in reconstructing high-frequency details; however, they face issues related to ambiguous global height changes within a scene. To address this limitation, we introduce a high-low frequency separation method for 3D reconstruction from input images. In this approach, the low-frequency component is reconstructed using the SfS algorithm, whereas the high-frequency image contains too much noise and is discarded. The proposed high-low frequency separation SfS method enhances the reconstruction accuracy of traditional SfS methods. Figure 3
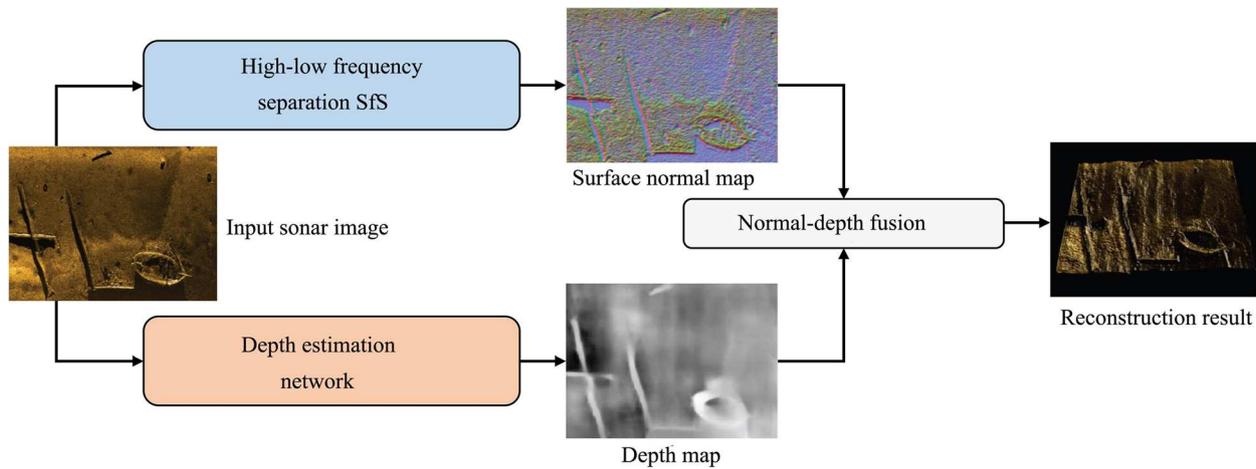
Fig. 2 The overall structure of the proposed underwater side-scan sonar image 3D reconstruction method
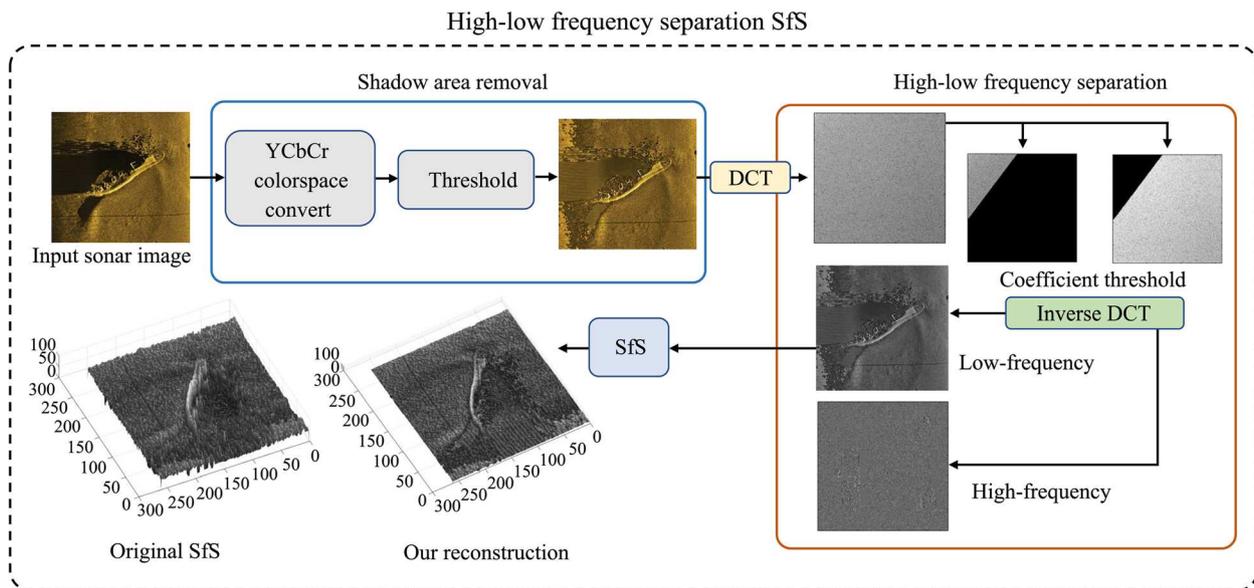

Fig. 3 Framework of the proposed high-low frequency separation SfS method. We first remove the shadow of the input sonar image and use the DCT (Ahmed et al. 1974) to split the sonar images into high- and low-frequency components. The output of SfS (Brooks and Horn 1985) using the low-frequency component, with less noise, obviously outperforms the results of the original sonar image

shows the framework of the proposed high and low-frequency separation SfS method.

In the context of side-scan sonar imaging, which is similar to the optical imaging model, the propagation of sound waves can be obstructed by tall targets (Deb and Suny 2014). This obstruction forms sound shadows behind the obstructing targets. If these sound shadows are not properly accounted for in the SfS algorithm, they can be incorrectly interpreted as depressions, thus impacting the accuracy of the target shape reconstruction. As shown in Fig. 3, we first remove the shadow areas in the input sonar images.

Inspired by Deb and Suny (2014), we first transform the red-green-blue (RGB) colorspace sonar image into the YCbCr colorspace. The Y channel corresponds to the brightness information of the image, while the Cb and Cr channels represent the color information. To identify shaded areas, we calculate the mean value of the Y channel and examine individual pixels within the image. Pixels with brightness values lower than the mean are then labeled as shadow areas.

We compute the means of the Y, Cb, and Cr channels, along with the standard deviation of the Y channel, separately for the shadowed and non-shadowed regions

Ju *et al. Intelligent Marine Technology and Systems* (2024) 2:4

Page 5 of 15

within the image. We denote the means of the three channels for the shadowed area as μ_{SY}, μ_{SCb}, μ_{SCr}, and for the non-shadowed area as μ_Y, μ_{Cb}, μ_{Cr}, with the ratio of the standard deviation of the Y channel as λ. In this context, the coefficients α_y, α_b, and α_r for shadow removal can be defined as follow:

$$\alpha_y = \mu_Y - \lambda \times \mu_{SY}, \tag{1}$$

$$\alpha_b = \mu_{Cb} - \lambda \times \mu_{SCb}, \tag{2}$$

$$\alpha_r = \mu_{Cr} - \lambda \times \mu_{SCr}. \tag{3}$$

In this case, for a pixel in the YCbCr image with the value $[Y, Cb, Cr]$, its value after shadow removal is denoted as $[Y', Cb', Cr']$ can be formulated as follows:

$$Y' = \alpha_y + \lambda \times Y, \tag{4}$$

$$Cb' = \alpha_b + \lambda \times Cb, \tag{5}$$

$$Cr' = \alpha_r + \lambda \times Cr. \tag{6}$$

After removing the shadows in sonar images, we employ DCT (Ahmed et al. 1974) to separate the high- and low-frequency components of the sonar image. DCT is a commonly used data image compression technique that converts images from spatial to frequency distribution. We apply a mask operation to the transformed image, zeroing out coefficients smaller than a specific threshold. This process results in compressed low-frequency image coefficients. We then calculate the high-frequency coefficients by subtracting the compressed image coefficients from the original coefficients. The low-frequency coefficients capture the grayscale variations in the image, whereas the high-frequency coefficients represent the edges and texture details of the target shape. We perform an inverse DCT on the low- and high-frequency coefficients to obtain different images for the high- and low-frequency components.

Finally, we employ a cost function minimization SfS method (Brooks and Horn 1985) to reconstruct the surface normal of the image. Figure 3 shows that since noise primarily exists in the high-frequency component, the proposed high-low frequency separation SfS method significantly improves the accuracy of the results by reducing noise, compared with the results of Brooks and Horn (1985) via the original sonar image.

3.2 Depth estimation network with dilated convolution and attention mechanism

In this section, we propose a novel monocular depth estimation network based on an encoder-decoder architecture to estimate the depth information of the captured sonar image. The encoder part extracts dense features and fundamental contextual information from the input image via a ResNet-101 module (He et al. 2016). It captures contextual information at multiple scales using convolution kernels with different expansion rates (dilated convolution) (Yu et al. 2017). In contrast, the decoder employs a local planar guidance module (Lee et al. 2019) to replace nearest neighbor upsampling, enabling the network to efficiently restore the feature map to its original resolution. The decoder incorporates an attention mechanism (Vaswani et al. 2017) for differential connections, allowing the predicted depth map to be specific to objects while maintaining the original prediction accuracy. This refinement enhances the edges and local details, catering to subsequent high-precision 3D reconstruction requirements. Figure 4 shows the detailed structure of the proposed depth estimation network.

Figure 4 shows that our proposed depth estimation network follows an encoding-decoding scheme. The overall process involves reducing the input feature resolution to $\frac{p}{8}$ and restoring the depth map to the original resolution p for dense depth prediction. First, a dense feature extractor, which includes a backbone network like ResNet-101 (He et al. 2016), is used to downsample the original monocular input image to $\frac{p}{8}$ resolution (where p represents the input image resolution). Subsequently, multi-scale dilated convolutional layers are employed with an attention mechanism to extract contextual information. The expansion rates r used are 3, 6, 12, and 24. Each scale's dilated convolution layer is combined with an attention module. During the decoding phase, local planar guidance layers (Lee et al. 2019) are used for layer-by-layer upsampling of the multi-scale feature maps, progressing from $\frac{p}{8}$ resolution to p, finally restoring them to the original image resolution. Finally, a 1×1 convolution operation is applied to obtain the final high-resolution depth map d, which is a single-channel output.

The proposed depth estimation network is optimized using the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{depth}} + w\mathcal{L}_{\text{gradient}}, \tag{7}$$

where $\mathcal{L}_{\text{depth}}$ is the loss function on the depth map domain, and $\mathcal{L}_{\text{gradient}}$ is the gradient constraint on the gradient of the depth map with a weight w empirically set to 0.1. Specifically, we define the $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{gradient}}$ as follows:

$$\mathcal{L}_{\text{depth}} = \frac{1}{T} \sum_i^T (d_i - \tilde{d}_i), \tag{8}$$
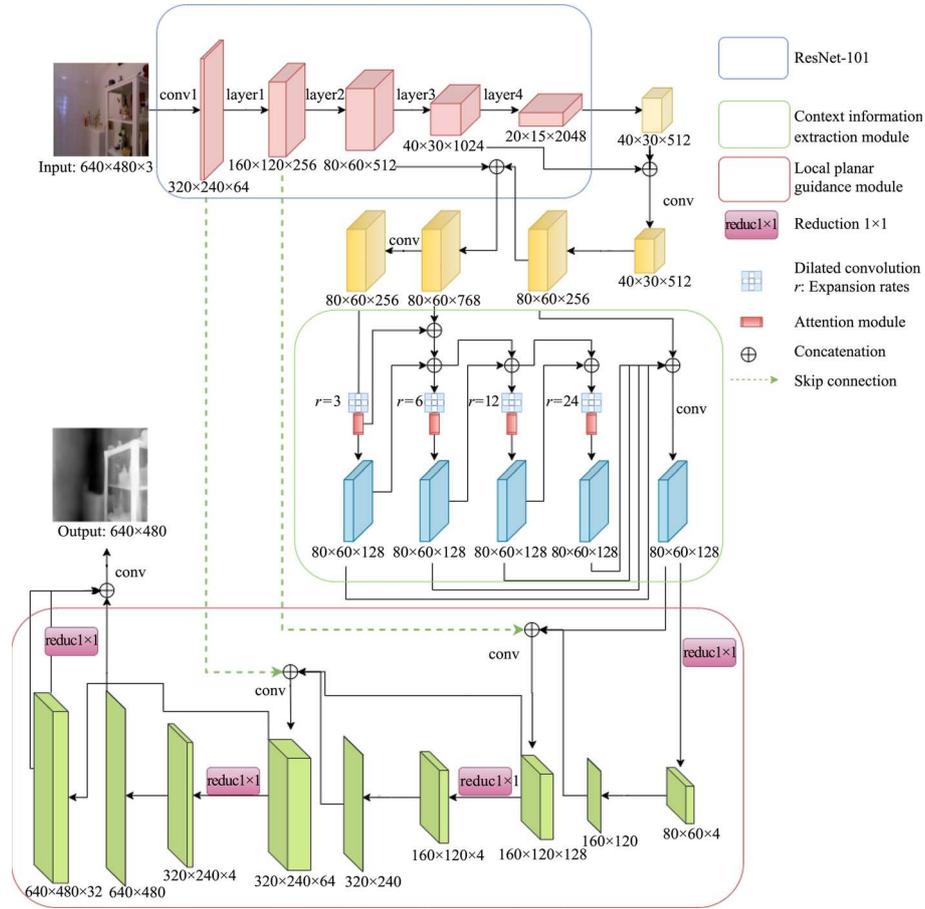
and

Fig. 4 Detailed structure of the proposed depth estimation network

$$\mathcal{L}_{\text{gradient}} = \frac{1}{T}\sum_{i}^{T}[(\frac{\partial d_i}{\partial x} - \frac{\partial \tilde{d}_i}{\partial x})^2 + (\frac{\partial d_i}{\partial y} - \frac{\partial \tilde{d}_i}{\partial y})^2], \tag{9}$$

where d_i and $\tilde{d}_i$ represent a pixel on the ground truth and the estimated depth map with the position index i, T stands for the number of total pixels. $\frac{\partial d_i}{\partial x}$ and $\frac{\partial d_i}{\partial y}$ represent the gradient of d_i on x and y directions, respectively.

3.3 Normal-depth fusion algorithm
In this section, we combine the surface normal map obtained from the high-low frequency separation SfS (Section 3.1) with the depth map obtained from the depth estimation network (Section 3.2). This fusion allows us to leverage the benefits of detailed and global results.

We capitalize on the strengths of SfS and depth estimation to jointly optimize the final height map. The main aim is to accurately compute the global scene height while preserving the local intricacies of the object's shape. This involves using the object's surface normal vector obtained from SfS to guide our target surface normal vector and utilizing the depth map obtained from depth estimation to guide the target's height map. Additionally, we introduce supplementary surface normal gradient constraints to ensure a smooth transition in the object's surface normal vector. These three components are combined to formulate the final fusion energy function as follows:

$$\iint_{(x,y)} m(D_f - D_e)^2 + (1 - m)\left((n_f - n_s)^2 + (n_{fx} + n_{fy})^2\right)\mathrm{d}x\mathrm{d}y = \iint_{(x,y)} F\mathrm{d}x\mathrm{d}y, \tag{10}$$

where D_f denotes the target depth map after surface optimization, D_e represents the depth obtained from depth estimation, n_s denotes the surface normal vector obtained

using the light and dark shape recovery algorithm, n_f denotes the target surface normal, and n_{fx} and n_{fy} denote the gradients of the target surface normal vector in the x and y directions, respectively. The variable m represents the weight proportion of each component. The surface normal vector can be expressed in terms of height gradient as follows:

$$
\begin{aligned}
\left(p_f, q_f, -1\right) &= \left(\frac{n_x}{n_z}, \frac{n_y}{n_z}, -1\right) \\
&= \left(\frac{\partial D_f}{\partial x}, \frac{\partial D_f}{\partial y}, -1\right) \text{ s.t. } n_f \\
&= \left(n_x, n_y, n_z\right),
\end{aligned}
\tag{11}
$$

which can be solved using the Euler-Lagrange (E-L) equation. The E-L equation for Eq. (11) can be expressed as follows:

$$
m\left(D_f - D_e\right) - (1 - m)\left(\frac{\partial^2 D_f}{\partial x^2} + \frac{\partial^2 D_f}{\partial y^2} + \frac{\partial^3 D_f}{\partial x^3} + \frac{\partial^3 D_f}{\partial y^3} - \frac{\partial^2 D_s}{\partial x^2} - \frac{\partial^2 D_s}{\partial y^2}\right) = 0,
\tag{12}
$$

where each part can be expressed via differential representation:

$$
\frac{\partial^2 D(x,y)}{\partial x^2} \approx \Delta^2 D(x) = D(x+1, y) - 2D(x,y) + D(x-1, y),
\tag{13}
$$

$$
\frac{\partial^2 D(x,y)}{\partial y^2} \approx \Delta^2 D(y) = D(x, y+1) - 2D(x,y) + D(x, y-1),
\tag{14}
$$

$$
\frac{\partial^3 D(x,y)}{\partial x^3} \approx \Delta^2 D(x+1) - \Delta^2 D(x),
\tag{15}
$$

$$
\frac{\partial^3 D(x,y)}{\partial y^3} \approx \Delta^2 D(y+1) - \Delta^2 D(y).
\tag{16}
$$

Eqs. (13), (14), (15), and (16) can be used to solve the E-L equation, as follows:

$$
D_f = \frac{1}{m}\left[(1 - m)\widetilde{D} + mD_e - (1 - m)\left(\frac{\partial^2 D_s}{\partial x^2} + \frac{\partial^2 D_s}{\partial y^2}\right)\right],
\tag{17}
$$

where $\widetilde{D}$ is:

$$
\begin{aligned}
\widetilde{D} =& D(x+2, y+1) + D(x, y+1) + D(x+1, y+2) \\
&+ D(x+1, y) - 4D(x+1, y+1),
\end{aligned}
\tag{18}
$$

where the iterative representation of Eq. (17) is given by:

$$
D_f^{n+1} = \frac{1}{m}\left[(1 - m)\widetilde{D}^n + mD_e - (1 - m)\left(\frac{\partial^2 D_s}{\partial x^2} + \frac{\partial^2 D_s}{\partial y^2}\right)\right],
\tag{19}
$$

where $\widetilde{D}^n$ represents the result of the nth iteration. The result of the nth iteration can be used to obtain the result of the $n+1$ iteration. If the number of iterations exceeds the maximum threshold, or the absolute difference between the results of two consecutive iterations is less than the minimum threshold, the iteration can be terminated, and the output obtained at that point is considered the final target result. Algorithm 1 summarizes the proposed fusion method.

---

1 **Require:** height map D_s from SfS (Section 3.1), depth map D_e from depth estimation network (Section 3.2)
2 1: $i \leftarrow 0$
3 2: $D_f^i \leftarrow$ Eqs. (17) (18)
4 3: **while True do**
5 4: $\widetilde{D}_f^i \leftarrow$ Eq. (18)
6 5: $\widetilde{D}_f^{i+1} \leftarrow$ Eq. (19)
7 6: $E = \left|D_f^{i+1} - D_f^i\right|$ $i+=1$
8 7:   **if** $E < K$ OR $i > n$ **then**
9 8:   break
10 9:   **end if**
11 10: **end while**
12 11: return D_f

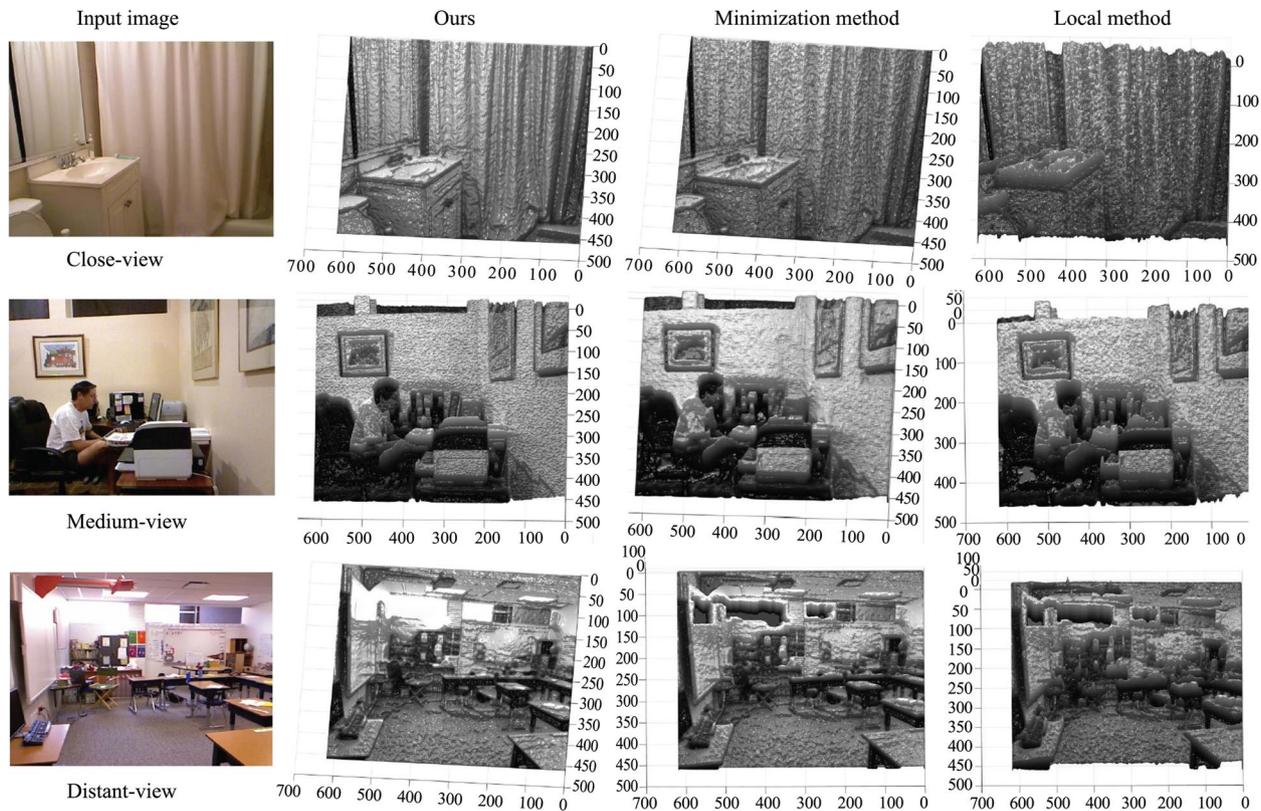**Algorithm 1** The proposed fusion algorithm

Fig. 5 Reconstruction results of the proposed SfS method compared with those of the minimization method (Brooks and Horn 1985) and local method (Pentland 1984) on the NYU-depth-v2 dataset (Silberman et al. 2012)

4 Experimental results

4.1 Implementation details

Our method is implemented using PyTorch. We used the Adam optimizer with default settings ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The training was performed on a single RTX 2080Ti graphics processing unit (GPU). Given the similarity between the imaging model of side-scan sonar and optical images and the limited availability of side-scan sonar training datasets, we opted to use the NYU-depth-v2 dataset for training. Specifically, we employed 30000 samples with data augmentation (including rotation and cropping) for training and used 1000 samples from the NYU-depth-v2 dataset for validation. We also test the proposed method on the synthetic side-scan sonar dataset (Gwon et al. 2017) and real side-scan sonar dataset[1].

4.2 Evaluation metrics

To evaluate the performance of the proposed SfS method, we employ the Lambertian reflection model to invert the reconstructed height map into a reflection map corresponding to the object for quantitative comparison.

This is because we do not have the ground-truth surface normal map of the side-scan sonar images. The evaluation of the reconstruction quality is based on the correlation coefficient r and signal-to-noise ratio SNR between the reflection image I' and the original grayscale image I. These metrics serve as indirect evaluation indices of the reconstruction effectiveness, which are given as follows:

$$r = \frac{\sum_{y=1}^{N} \sum_{x=1}^{M} \left(I'(x,y) - \bar{I}'\right) \times (I(x,y) - \bar{I})}{\sqrt{\sum_{y=1}^{N} \sum_{x=1}^{M} \left(I'(x,y) - \bar{I}'\right)^2} \sqrt{\sum_{y=1}^{N} \sum_{x=1}^{M} (I(x,y) - \bar{I})^2}}, \quad (20)$$

where $\bar{I}'$ and $\bar{I}$ represent the average value of I' and I, respectively. The correlation coefficient r has a value between −1 and 1. When r is closer to 1, it indicates a stronger correlation between the reflection map and input image, suggesting a higher degree of similarity. This means a better reconstruction effect for the estimated height map. Additionally, a SNR indicates a higher proportion of signal in the image, resulting in less noise and better image quality.

Furthermore, to assess the performance of the proposed depth estimation network and fused output, we evaluate several widely used metrics, such as threshold metrics ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$), absolute relative

---

[1] https://www.edgetech.com/underwater-technology-gallery/

error (*AbsRel*), root mean square error (*RMSE*), and log10 error. For these metrics, the smaller their values, the better the result.

4.3 Effectiveness of the proposed SfS method

To assess the performance of the proposed high-low frequency separation SfS method, we conducted a comparative evaluation against the minimization method (Brooks and Horn 1985) and local method (Pentland 1984). We first test them on the NYU-depth-v2 test dataset (Silberman et al. 2012), with visualized results in Fig. 5 and quantitative results in Table 1. Then, we compare the estimations on the real side-scan sonar dataset,

with visualized results in Fig. 6 and quantitative results in Table 2.

Figure 5 shows noticeable distortions in the reconstruction results of Brooks and Horn (1985) and Pentland (1984), with much noise. In contrast, the proposed method, which incorporates low-frequency height constraints, yields smoother reconstructions with fewer instances of missing height information. In the case of distant-view scenes, minimization and local methods are simply replaced with holes, which evidently fails to meet scene reconstruction requirements.

As presented in Table 1, the proposed high-low frequency separation SfS outperforms the minimization and local methods across all metrics with different samples. A

Table 1 Quantitative comparisons using the metrics *r* and *SNR* on the NYU-depth-v2 dataset (Silberman et al. 2012)

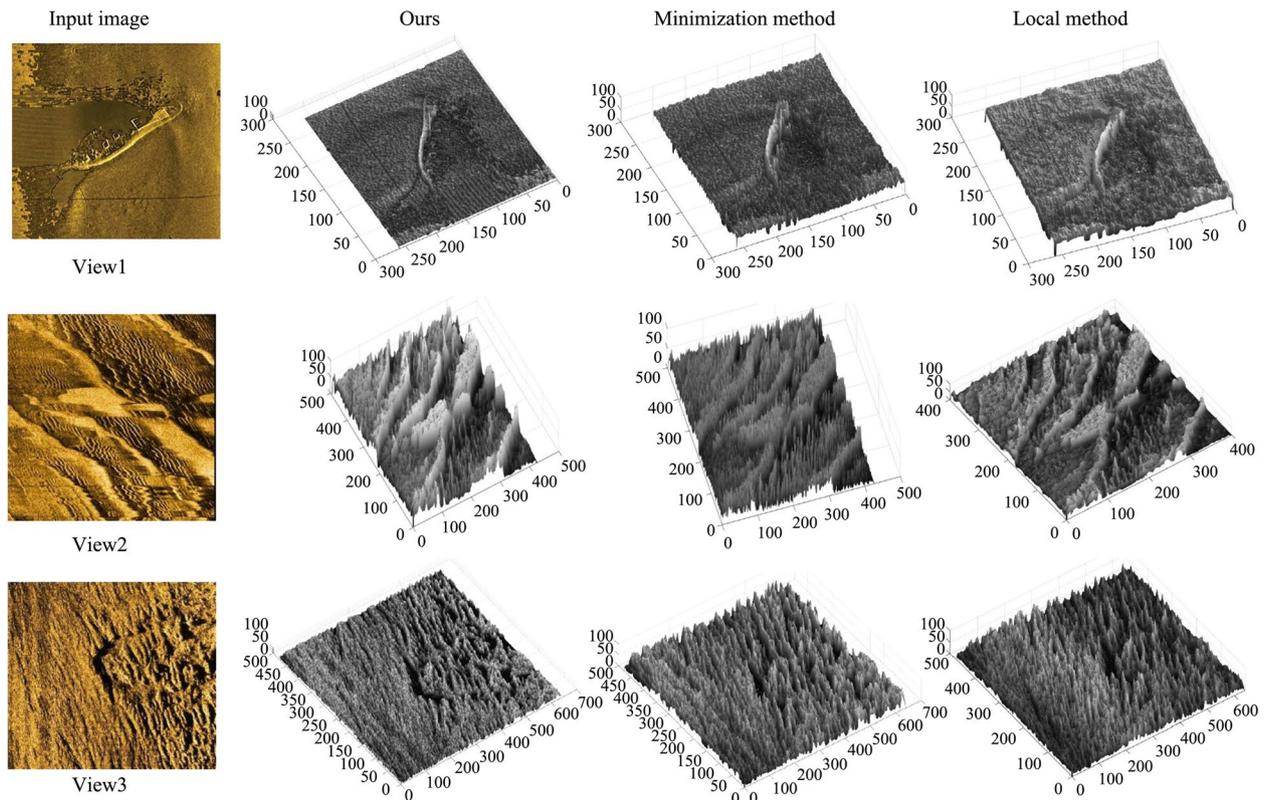| Method	Close-view		Medium-view		Distant-view	
	r	*SNR*	*r*	*SNR*	*r*	*SNR*
Ours	0.0131	0.7833	0.0551	0.7736	0.0351	0.8524
Minimization method (Brooks and Horn 1985)	0.0022	0.7060	0.0095	0.6333	0.0015	0.8188
Local method (Pentland 1984)	0.0020	0.6805	0.0094	0.6157	0.0015	0.7686

Fig. 6 Reconstruction results of the proposed SfS method compared with those of the minimization method (Brooks and Horn 1985) and local method (Pentland 1984) on the real side-scan sonar dataset

Table 2 Quantitative comparisons using the metrics *r* and *SNR* on the real side-scan sonar dataset

Method	View1		View2		View3	
	r	*SNR*	*r*	*SNR*	*r*	*SNR*
Ours	0.0716	0.9162	0.1077	0.8520	0.1391	0.8298
Minimization method (Brooks and Horn 1985)	0.0542	0.6325	0.0677	0.8281	0.0620	0.8352
Local method (Pentland 1984)	0.0512	0.5938	0.0602	0.7657	0.0593	0.8062

higher *r* and *SNR* of the proposed method demonstrates that our reconstructed height map retains richer local details and less noise, demonstrating the effectiveness of the proposed SfS method.

We conducted comparative experiments on the real side-scan sonar dataset to further evaluate the proposed method. As shown in Fig. 6, traditional SfS methods (Pentland 1984; Brooks and Horn 1985) show height maps with more noise and less smoothness. In contrast, the height map reconstructed using our method suppresses the impact of high-frequency noise on the reconstruction quality due to the addition of low-frequency height constraint information. For example, in the scene of view3, the height map reconstructed using

our method delicately represents the undulations of the seafloor ripple topography, resulting in gentler terrain features. In contrast, traditional methods produce rougher ripple topography with larger undulations. The quantitative comparison demonstrates the effectiveness of the proposed method in handling the real side-scan sonar dataset.

4.4 Effectiveness of the proposed depth estimation network

In this experiment, we assess the effectiveness of the proposed depth estimation network incorporating dilated convolution and an attention mechanism. We conducted our initial evaluation on the NYU-depth-v2
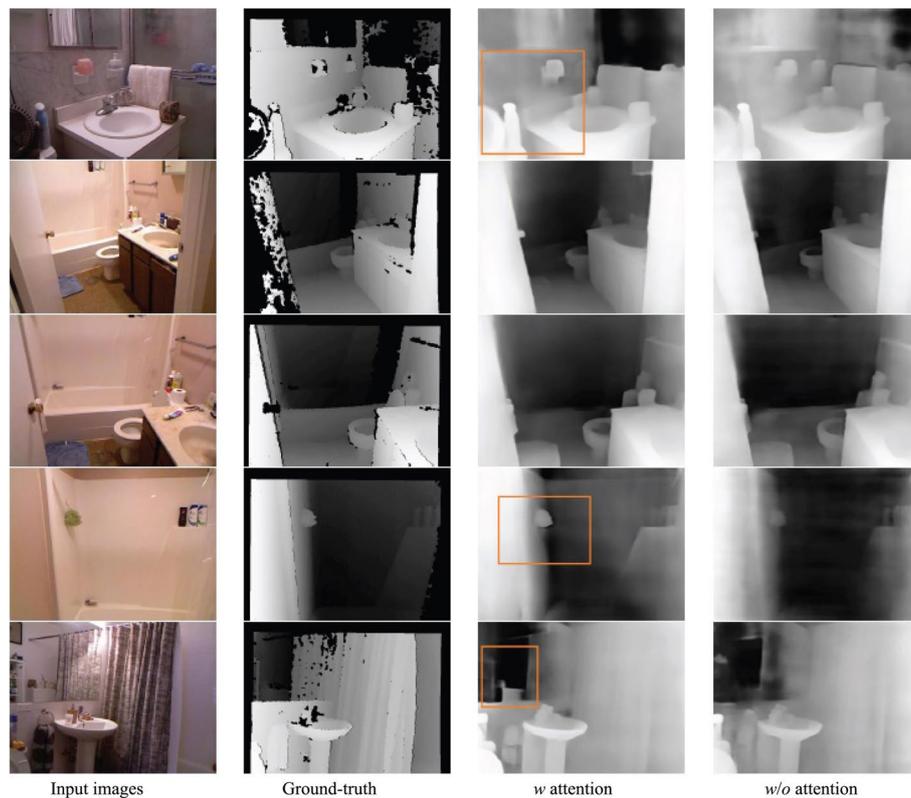

Input images   Ground-truth   *w* attention   *w/o* attention

Fig. 7 Ablation study of attention module (Vaswani et al. 2017) in our depth estimation network. The orange boxes represent the regions with detailed structures

Table 3 Quantitative comparisons on the NYU-depth-v2 test set

Method	$\delta < 1.25\uparrow$	$\delta < 1.25^2\uparrow$	$\delta < 1.25^3\uparrow$	AbsRel↓	RMSE↓	log10↓
Eigen (Eigen et al. 2014)	0.769	0.950	0.988	0.158	0.641	-
Laina (Laina et al. 2016)	0.811	0.953	0.988	0.127	0.573	0.055
Xu (Xu et al. 2017)	0.811	0.954	0.987	0.121	0.586	0.052
Lee (Lee et al. 2018)	0.815	0.963	0.991	0.139	0.572	-
SC-DepthV2 (Bian et al. 2021)	0.820	0.956	-	0.138	0.532	0.059
Qi (Qi et al. 2018)	0.834	0.960	0.990	0.128	0.532	0.059
Yin (Yin et al. 2019)	0.875	0.976	0.994	0.108	0.416	0.048
BTS (Lee et al. 2019)	0.882	0.978	0.994	0.110	0.396	0.047
Song (Song et al. 2021)	0.895	0.983	0.996	0.105	0.384	0.044
Ours	0.881	0.985	0.995	0.109	0.381	0.046
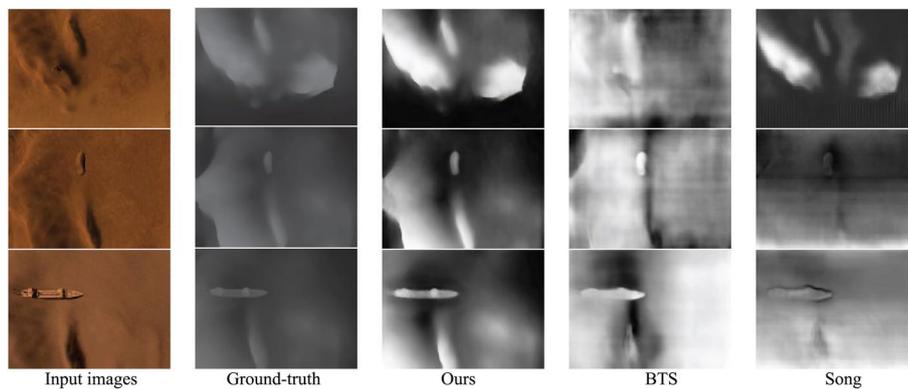
Fig. 8 Visualized examples on the synthetic side-scan sonar dataset (Gwon et al. 2017), compared with BTS (Lee et al. 2019) and Song (Song et al. 2021)

dataset (Silberman et al. 2012). Figure 7 shows an ablation experiment to assess the efficacy of the incorporated attention module (Vaswani et al. 2017). Subsequently, we conducted a comparative analysis with recent state-of-the-art (SOTA) depth estimation networks on the NYU-depth-v2 test set, with results summarized in Table 3. Finally, we validate the proposed method using a synthetic side-scan sonar dataset (Gwon et al. 2017) and a real side-scan sonar dataset, as shown in Figs. 8 and 9. Since we know the ground truth in the synthetic side-scan sonar dataset, we further quantitatively compare them in Table 3.

Figure 7 shows that our network with the attention module accurately predicts global depth, offering superior performance in predicting object boundaries and scene details compared with the network without the attention mechanism. Furthermore, we compare our methods with nine recent networks (as listed in Table 3) using the NYU-depth-v2 test set. The results demonstrate that the proposed depth estimation

network achieves the best or second-best performance across six evaluation metrics.

We further validate the depth estimation performance of the proposed method on the side-scan sonar dataset (Fig. 8 and Table 4). As shown in Fig. 8, the comparison clearly demonstrates that our network accurately captures the global scene depth while presenting local details more intricately. The proposed network excels in rendering small objects with precision. In contrast, the depth maps obtained from BTS (Lee et al. 2019) suffer from information loss in edge details, making it challenging to distinguish objects. Song (Song et al. 2021) also exhibits numerous depth estimation errors and missing information. Our depth map closely resembles actual depth, delivering superior performance. As presented in Table 4, the proposed method outperforms SOTA methods.

Figure 9 shows the qualitative evaluation of the proposed depth estimation network using real side-scan sonar images. As demonstrated in the second image, our method reveals finer details, highlighting the pipeline
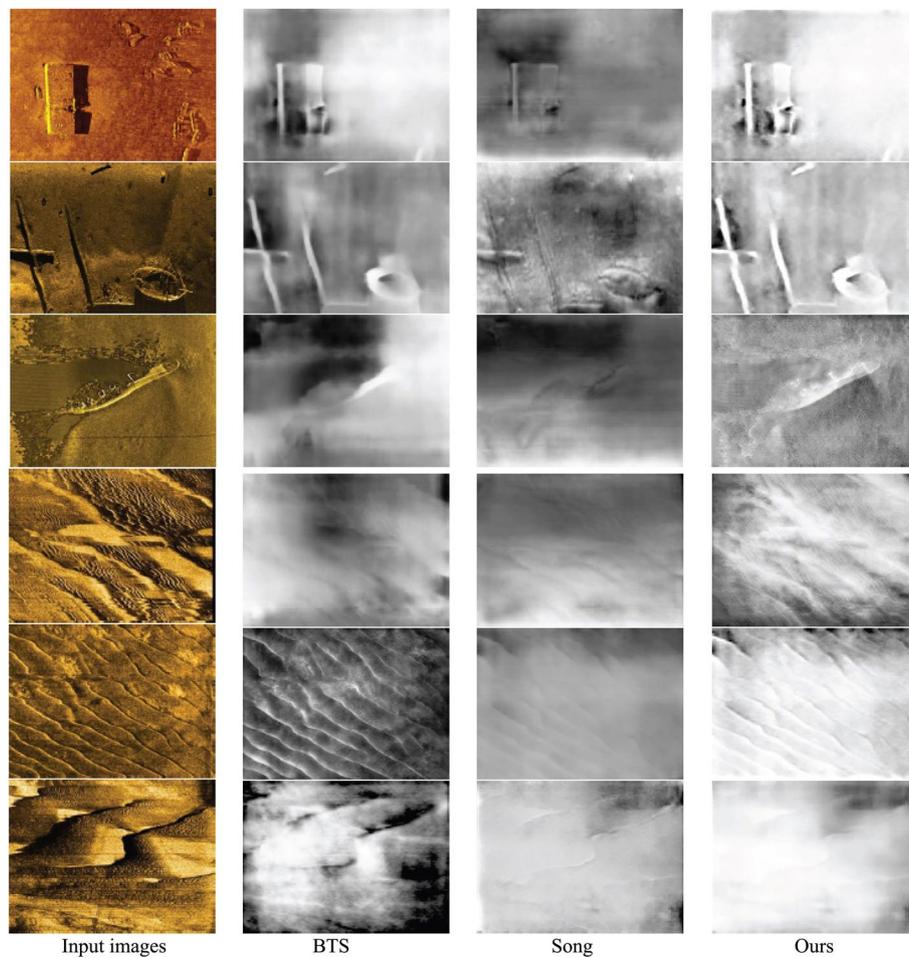
Fig. 9 Qualitative results on the real side-scan sonar dataset, compared with BTS (Lee et al. 2019) and Song (Song et al. 2021)

structure. The third and fourth images reveal significant depth information loss and blurry estimations in the BTS (Lee et al. 2019) and Song (Song et al. 2021) methods. In contrast, our method incorporates the attention mechanism (Vaswani et al. 2017) between the encoder and decoder, allowing the network to emphasize vital image features and prioritize local scene details. In the decoding phase, we replace the traditional deconvolution network with the integration of local plane guidance layers (Lee et al. 2019), enhancing the performance in estimating depth map details.

4.5 Effectiveness of the proposed normal-depth fusion algorithm

To evaluate the performance of our fusion method (Section 3.3), we compare it with the standalone SfS method (Section 3.1) and the independent depth estimation network (Section 3.2). Figure 10 and Table 4 show the results on the NYU-depth-v2 dataset, whereas Fig. 11 shows the results on real side-scan sonar images.

Figure 10 shows that the proposed SfS method can effectively capture the details but may yield inaccurate depth predictions for objects hidden in shadows and affected by noise. The depth map generated by the depth estimation algorithm accurately depicts global depth trends but lacks fine details. In contrast, the fusion algorithm produces a depth map that combines the correct global depth with rich details from the SfS algorithm. It particularly excels in preserving edge details. Table 4 also demonstrates the improved performance of the proposed fusion algorithm. In summary, our fusion algorithm is superior in 3D reconstruction compared with the single SfS method and depth estimation network.

We also assess the proposed fused algorithm on the real side-scan sonar dataset. To assess the applicability of the fusion method in various side-scan sonar image scenarios, we conducted comparative experiments involving the reconstruction of underwater aircraft wreckage, underwater structures, and seabed terrain with distinct characteristics (Fig. 11).
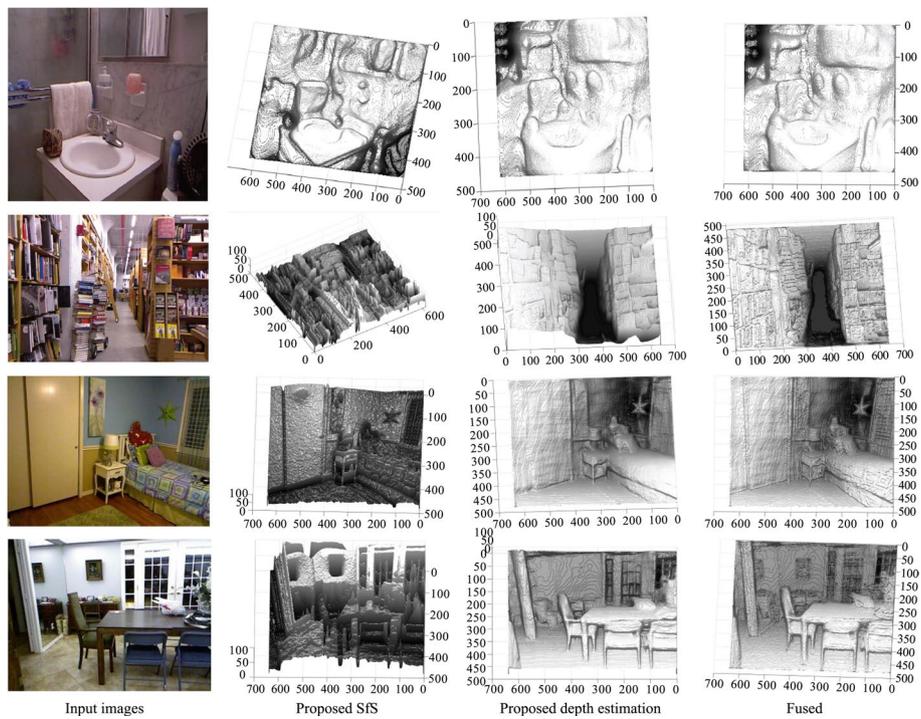
Fig. 10 Visualized fused results on the NYU-depth-v2 dataset (Silberman et al. 2012)

Table 4 Quantitative comparisons of the fused method on the NYU-depth-v2 dataset (Silberman et al. 2012)

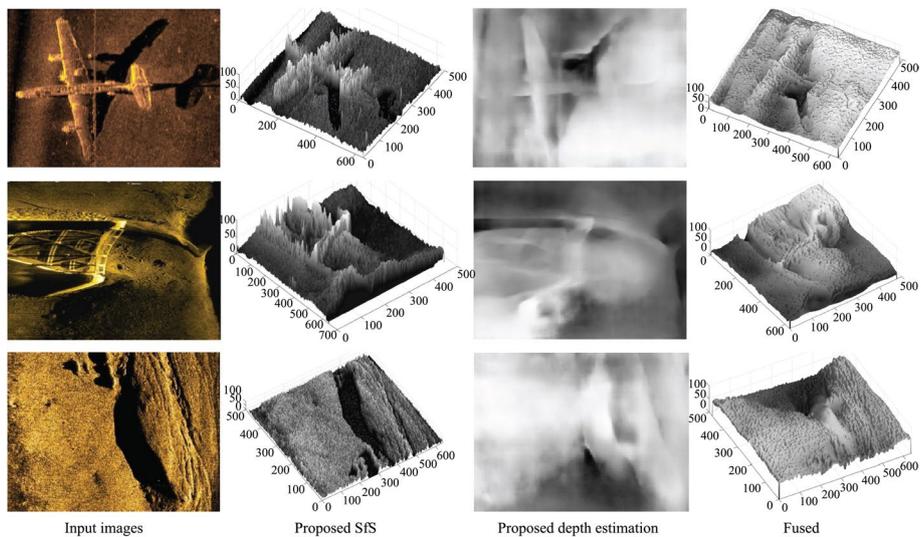| Method	δ < 1.25↑	δ < 1.25²↑	δ < 1.25³↑	AbsRel↓	RMSE↓	log10↓
Only depth	0.881	0.985	0.995	0.109	0.381	0.046
Fused method	0.883	0.985	0.995	0.107	0.379	0.045

Fig. 11 Visualized fused results on the real side-scan sonar images

In the underwater aircraft wreckage view, the height map obtained using the SfS algorithm exhibits relatively pronounced fluctuations. In contrast, the fused height map exhibits smoother variations on the seafloor, with reduced disruptions caused by shadows in the image. Additionally, in the underwater structure view, the primary subject in the image appears to be a metal structure, resulting in high-brightness reflection. In this case, the normal map obtained using the SfS method effectively captures the contour details of the object. However, in the reconstructed height map, the edges of the target exhibit excessive jaggedness, which is associated with uneven brightness distribution in the image. The height map obtained using the final fusion algorithm successfully mitigates the excessive noise observed in the SfS algorithm. Moreover, fusion with the depth map results in a smoother and more natural overall height transition. Finally, in the seabed terrain view, the proposed SfS algorithm successfully captures the undulations of the seafloor terrain; however, the resulting 3D terrain appears excessively rough, lacking smoothness. The depth estimation process also encounters issues because it incorrectly estimates the depth in the latter part of the shadow. Although the height map exhibits smoother characteristics during the reconstruction fusion process, it still contains incorrect height estimates, particularly in the upper-left portion. We argue that if the depth estimation process can accurately estimate the depth of the image, the overall reconstruction quality should surpass that of the SfS algorithm.

5 Conclusions

In this study, we presented a novel 3D reconstruction approach for underwater side-scan sonar images. Our method addresses the issues of noise and errors in global information encountered in traditional methods. We first proposed an SfS method that separates high- and low-frequency components in side-scan sonar images using DCT. This technique helps mitigate the problem of noise in traditional SfS methods. We developed a monocular depth estimation network incorporating dilated convolution and an attention mechanism. This network accurately estimates the depth map of sonar images, providing global information that complements the SfS results. Finally, we designed a fusion algorithm to combine the surface normal and depth maps obtained using our methods. The fused 3D height map provides more accurate results with less noise and precise global structural information. The experimental results on various datasets validate the effectiveness of the proposed high-low frequency separation SfS method, depth estimation network, and fusion algorithm.

Authors' contributions
Yakun Ju: Conceptualization, Methodology, Formal Analysis, Resources, Supervision, Writing-Original Draft; Jingchun Zhou: Conceptualization, Methodology, Supervision; Shitong Zhou: Methodology, Software, Validation, Data Curation, Writing-Original Draft; Hao Xie: Software, Methodology, Validation, Data Curation; Cong Zhang: Software, Validation, Data Curation; Jun Xiao: Validation, Formal Analysis, Data Curation; Cuixin Yang: Validation, Data Curation; Jianyuan Sun: Conceptualization, Writing-Original Draft.

Availability of data and materials
No associated data and materials.

Declarations

Ethics approval and consent to participate
No ethical approval was necessary for this work. Submission declaration and verification: We confirm that our work is original. Our manuscript has not been published, nor is it currently under consideration for publication elsewhere.

Consent for publication
Yakun Ju, Jingchun Zhou, Shitong Zhou, Hao Xie, Cong Zhang, Jun Xiao, Cuixin Yang, and Jianyuan Sun declare that they consent to participate.

Competing interests
No conflict of interest exists in the submission of this manuscript, and the manuscript is approved by all authors for publication. All the authors listed have approved the manuscript that is enclosed. The authors declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Yakun Ju is one of the Editorial Board Members, but he was not involved in the journal's review of, or decision related to, this manuscript.

References
Ahmed N, Natarajan T, Rao KR (1974) Discrete cosine transform. IEEE Trans Comput C-23(1):90–93

Bellman R (1966) Dynamic programming. Science 153(3731):34–37

Bian JW, Zhan H, Wang NY, Chin TJ, Shen CH, Reid I (2021) Auto-rectify network for unsupervised indoor depth estimation. IEEE Trans Pattern Anal Mach Intell 44(12):9802–9813

Brooks MJ, Horn BK (1985) Shape and source from shading. In: Proceedings of the International Joint Conference on Artificial Intelligence, Morgan Kaufmann, pp 932–936

Chen WF, Fu Z, Yang DW, Deng J (2016) Single-image depth perception in the wild. In: 30th Conference on Neural Information Processing Systems (NIPS), Barcelona, pp 730–738

Coiras E, Petillot Y, Lane DM (2007) Multiresolution 3-D reconstruction from side-scan sonar images. IEEE Trans Image Proc 16(2):382–390

Deb K, Suny AH (2014) Shadow detection and removal based on YCbCr color space. Smart Comput Rev 4(1):23–33

Durá E, Bell J, Lane D (2004) Reconstruction of textured seafloors from side-scan sonar images. IEE Proc-Radar Sonar Navig 151(2):114–126

Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. In: 28th Conference on Neural Information Processing Systems (NIPS), Montreal, pp 2366–2374

Fan H, Qi L, Ju YK, Dong JY, Yu H (2017) Refractive laser triangulation and photometric stereo in underwater environment. Opt Eng 56(11):113101

Frankot RT, Chellappa R (1988) A method for enforcing integrability in shape from shading algorithms. IEEE Trans Pattern Anal Mach Intell 10(4):439–451

Gan YK, Xu XY, Sun WX, Lin L (2018) Monocular depth estimation with affinity, vertical pooling, and label enhancement. In: 15th European Conference on Computer Vision (ECCV), Munich, pp 232–247

Gwon DH, Kim J, Kim MH, Park HG, Kim TY, Kim A (2017) Development of a side scan sonar module for the underwater simulator. In: 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Jeju, pp 662–665

He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, pp 770–778

Horn BKP (1970) Shape from shading: a method for obtaining the shape of a smooth opaque object from one view. PhD thesis, Massachusetts Institute of Technology

Ikeuchi K, Horn BKP (1981) Numerical shape from shading and occluding boundaries. Artif Intell 17(1–3):141–184

Ju YK, Jian MW, Dong JY, Lam KM (2020a) Learning photometric stereo via manifold-based mapping. In: IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, pp 411–414

Ju YK, Jian MW, Guo SX, Wang YY, Zhou HY, Dong JY (2021) Incorporating lambertian priors into surface normals measurement. IEEE Trans Instrum Meas 70:1–13

Ju YK, Jian MW, Wang C, Zhang C, Dong JY, Lam KM (2023a) Estimating high-resolution surface normals via low-resolution photometric stereo images. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2023.3301930

Ju YK, Qi L, He JC, Dong XH, Gao F, Dong JY (2020b) MPS-Net: learning to recover surface normal for multispectral photometric stereo. Neurocomputing 375:62–70

Ju YK, Shi BX, Chen Y, Zhou HY, Dong JY, Lam KM (2023b) GR-PSN: learning to estimate surface normal and reconstruct photometric stereo images. IEEE Trans Vis Comput Graph. https://doi.org/10.1109/TVCG.2023.3329817

Ju YK, Shi BX, Jian MW, Qi L, Dong JY, Lam KM (2022) NormAttention-PSN: a high-frequency region enhanced photometric stereo network with normalized attention. Int J Comput Vis 130(12):3014–3034

Ju YK, Zhang C, Huang SS, Rao Y, Lam KM (2023c) Learning deep photometric stereo network with reflectance priors. In: IEEE International Conference on Multimedia and Expo (ICME), Brisbane, pp 2027–2032

Jung H, Kim Y, Min D, Oh C, Sohn K (2017) Depth prediction from a single image with conditional adversarial networks. In: 24th IEEE International Conference on Image Processing (ICIP), Beijing, pp 1717–1721

Key WH (2000) Side scan sonar technology. In: OCEANS 2000 MTS/IEEE Conference and Exhibition, Providence, pp 1029–1033

Kimmel R, Bruckstein AM (1995) Tracking level sets by level sets: a method for solving the shape from shading problem. Comput Vis Image Underst 62(1):47–58

Kong CQ, Chen BL, Li HL, Wang SQ, Rocha A, Kwong S (2022) Detect and locate: exposing face manipulation by semantic- and noise-level telltales. IEEE Trans Inf Forensic Secur 17:1741–1756

Kong CQ, Chen BL, Yang WH, Li HL, Chen PL, Wang SQ (2021) Appearance matters, so does audio: revealing the hidden face via cross-modality transfer. IEEE Trans Circ Syst Video Technol 32(1):423–436

Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N (2016) Deeper depth prediction with fully convolutional residual networks. In: 4th IEEE International Conference on 3D Vision (3DV), Stanford, pp 239–248

Lee JH, Han MK, Ko DW, Suh IH (2019) From big to small: multi-scale local planar guidance for monocular depth estimation. Preprint arXiv:1907.10326

Lee JH, Heo M, Kim KR, Kim CS (2018) Single-image depth estimation based on fourier domain analysis. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, pp 330–339

Lee CH, Rosenfeld A (1985) Improved methods of estimating shape from shading using the light source coordinate system. Artif Intell 26(2):125–143

Li B, Shen CH, Dai YC, van den Hengel A, He MY (2015) Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, pp 1119–1127

Liu YR, Ju YK, Jian MW, Gao F, Rao Y, Hu YQ et al (2022) A deep-shallow and global-local multi-feature fusion network for photometric stereo. Image Vis Comput 118:104368

Luo AW, Li EL, Liu YL, Kang XG, Wang ZJ (2021) A capsule network based approach for detection of audio spoofing attacks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, pp 6359–6363

Pentland AP (1984) Local shading analysis. IEEE Trans Pattern Anal Mach Intell 6(2):170–187

Qi XJ, Liao RJ, Liu ZZ, Urtasun R, Jia JY (2018) GeoNet: geometric neural network for joint depth and surface normal estimation. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, pp 283–291

Rao Y, Ju YK, Wang S, Gao F, Fan H, Dong JY (2023) Learning enriched feature descriptor for image matching and visual measurement. IEEE Trans Instrum Meas 72:1–12

Rouy E, Tourin A (1992) A viscosity solutions approach to shape-from-shading. SIAM J Numer Anal 29(3):867–884

Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: 12th European Conference on Computer Vision (ECCV), Florence, pp 746–760

Song M, Lim S, Kim W (2021) Monocular depth estimation using laplacian pyramid-based depth residuals. IEEE Trans Circ Syst Video Technol 31(11):4381–4393

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. In: 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, pp 6000–6010

Wang YY, Ju YK, Jian MW, Lam KM, Qi L, Dong JY (2020) Self-supervised depth completion with attention-based loss. In: International Workshop on Advanced Imaging Technology (IWAIT), Yogyakarta, pp 517–524

Xiao J, Jiang XY, Zheng NX, Yang H, Yang YF, Yang YQ et al (2023) Online video super-resolution with convolutional kernel bypass grafts. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2023.3243615

Xiao J, Liu TS, Zhao R, Lam KM (2021) Balanced distortion and perception in single-image super-resolution based on optimal transport in wavelet domain. Neurocomputing 464:408–420

Xu D, Ricci E, Ouyang WL, Wang XG, Sebe N (2017) Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, pp 161–169

Yin W, Liu YF, Shen CH, Yan YL (2019) Enforcing geometric constraints of virtual normal for depth prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, pp 5683–5692

Yu F, Koltun V, Funkhouser T (2017) Dilated residual networks. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, pp 636–644

Yu Y, Wang YF, Yang WH, Lu SJ, Tan YP, Kot AC (2023) Backdoor attacks against deep image compression via adaptive frequency trigger. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Vancouver, pp 12250–12259

Zhang C, Lam KM, Wang Q (2023a) CoF-Net: a progressive coarse-to-fine framework for object detection in remote-sensing imagery. IEEE Trans Geosci Remote Sens 61:1–17

Zhang C, Su JR, Ju YK, Lam KM, Wang Q (2023b) Efficient inductive vision transformer for oriented object detection in remote sensing imagery. IEEE Trans on Geosci Remote Sens 61:1–20

Zheng QF, Chellapa R (1991) Estimation of illuminant direction, albedo, and shape from shading. IEEE Trans Pattern Anal Mach Intell 13(7):680–702

Zhou JC, Liu Q, Jiang QP, Ren WQ, Lam KM, Zhang WS (2023a) Underwater camera: improving visual perception via adaptive dark pixel prior and color correction. Int J Comput Vis. https://doi.org/10.1007/s11263-023-01853-3

Zhou JC, Pang L, Zhang DH, Zhang WS (2023b) Underwater image enhancement method via multi-interval subhistogram perspective equalization. IEEE J Ocean Eng 48(2):474–488

Publisher's Note