Neurocomputing 514 (2022) 362–373

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Underwater self-supervised depth estimation

Xuewen Yang [a,1], Xing Zhang [a,1], Nan Wang [a,*], Guoling Xin [a], Wenjie Hu [a]

[a] Ocean University of China, Qingdao, China

ARTICLE INFO

Article history:
Received 14 March 2022
Revised 4 September 2022
Accepted 15 September 2022
Available online 21 September 2022
Communicated by Zidong Wang

Keywords:
Underwater perception
Depth estimation
Self-supervised

ABSTRACT

Accurate underwater depth estimation is a cornerstone of reaching autonomous underwater exploration. However, it is incredibly tricky due to the inherent attenuation character and heavy noise. Fortunately, the depth-changing trend and underwater light attenuation are closely correlated, providing powerful clues for underwater depth estimation. Rather than simulating the underwater attenuation through formulas, we propose an underwater self-supervised depth estimation neural network in our work. With the guidance of multiple constraints, which are meticulously designed based on the comprehensive analyses of underwater characters, this network can learn the depth-changing trend by itself from attenuation information in underwater monocular videos. Our detailed experiments on underwater datasets prove that the proposed framework can obtain accurate and fine-grained depth maps. We believe the work may provide an economical solution for underwater perception.

1. Introduction

With the development of intervention capability, underwater robots no longer only perform as underwater vehicles, but more as the operation intelligent agents carrying out complex works [1], such as seafood fishing [2], underwater archaeology [3], inspection, maintenance, and repair applications [4], to name a few. The premise of safe and effective autonomous operations is the full perception of the environment [5], such as water depth, object depth, water pressure, temperature, salinity, *etc.* Among all the information that needs to be collected, object depth, which is the depth from the agent to the surroundings, is crucial.

Due to the inherent physical character, accurate underwater depth measurement is especially tricky if not impossible. Traditionally, sonar is applied to detect the long depth. There are different kinds of sonar, such as multi-beam sonar [6], single-beam sonar [7], and side-scan sonar [8]. However, the accuracy of close-range measurement is limited due to the limitation of the sonar's principle. It is unable to provide fine-grained information and cannot meet the needs of refined operations. To measure depth, several researchers propose other solutions based on optical principles, such as underwater laser line scanning [9,10] and range-gated imaging systems [11]. They all need special instruments, not suitable for low-weight robots. What's more, due to the absorption and scattering of light by water bodies and the great changes in turbidity of different waters, the accuracy of optical methods cannot be guaranteed.

In the atmosphere, LiDAR [12] shows great performance in depth measurement. However, they are expensive and affected by problems of serious scattering of laser in water, which prevent them from being widely used [13]. Binocular stereo, as another primary means of depth measurement, can provide dense depth maps but is also not suitable underwater [14]. Establishing the feature correspondences of underwater stereo images is challenging for the influence of attenuation and scattering on image quality [15]. There are various types of underwater cameras for obtaining intuitive images, which have become the most commonly used configuration for almost all underwater unmanned operation platforms. How to estimate the depth from monocular attracts the researchers' attention [16]. The attenuation effect, which consists of absorption and scattering, is assumed to be strongly correlated with depth [17]. Thus, some researchers try to restore depth information based on the underwater light attenuation and achieved some success, such as [16,18].

Recently, the rapid development of computer vision proposes a new solution for depth information restoration from single images [19]. Thus, in the paper, we first comprehensively analyze the data characteristics of underwater environments, then a monocular self-supervised depth estimation framework is designed for the underwater environment.

In summary, our main contributions are the following:

* Corresponding author.
E-mail address: wangnan@ouc.edu.cn (N. Wang).

[1] These authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.neucom.2022.09.122
0925-2312/

- The differences between the atmosphere and underwater environment are analyzed comprehensively, and an underwater self-supervised monocular depth estimation framework is designed based on these analyses.
- The relationship between underwater light attenuation and the depth-changing is used to help the network extract depth with the guidance of image enhancement.
- The gradient similarity and consistency between depth and optical flow are applied to refine the depth map.

Note that, the purpose of this paper is to solve underwater perception which is a more challenging problem than in atmosphere scenarios. To our knowledge, this is the first work to develop a self-supervised neural network that estimates depth maps directly from underwater monocular video without any ground truth. We believe that our work would make contributions to underwater-related robotics perception and marine exploration.

2. Related work

In 2014, Eigen et al. [19] firstly used deep learning (DL) to solve the problem of monocular depth estimation. They applied AlexNet to estimate depth in both NYU Depth v2 [20] and KITTI [21] and proposed a two-step strategy from coarse to fine. Since then, many variants have been proposed to improve depth estimation performance. Among them, several researchers used the full convolution neural network model to replace the previous fully connected network, improving the accuracy of estimated depth [22]. What's more, some methods [23,24] tried to increase the accuracy by including Conditional Random Fields (CRFs) [25]. All the aforementioned works are supervised by the ground truth, which is generated by Kinect camera [20], LiDAR [21], or stereo camera [21]. Obviously, supervised depth estimation methods have made a great success. However, these methods fall into a bottleneck on account of the difficulty of obtaining accurate depth values in many cases, such as underwater [26]. Thus, self-supervised learning methods are required in these cases.

Self-supervised learning based methods are developed, either using stereo pairs [27,28]or exploiting monocular sequences [29,30]. The core idea of self-supervised depth estimation is to establish pixel corresponding based on predicted depth maps, minimizing all the photometric reconstruction loss of paired pixels. In 2017, Zhou et al. [29] firstly used the correspondence of monocular video sequences to estimate depth. Recently, many efforts have been made to improve performance in different ways. One way is to take advantage of multiple tasks, such as segmentation [31], optical flow [32,33], object motion [34], surface normal [35], edge [36], and visual odometry [37]. Moreover, some works make use of sequence information by RNN [38], cost volume [39], and so on.

In respect of underwater situations, things usually get harder. However, the principle that the light may attenuate along with the depth gives an intuitive solution to estimate the depth. Some scholars tried to extract depth information based on this principle. Such methods are called physical model-based methods [40,41]. In recent years, in some underwater original image restoration techniques, the depth map is obtained as the by-product. For example, Wang et al. [16] proposed a new maximum attenuation identification method, which used the attenuation effect of absorption and scattering to obtain a depth map. Given the good performance of dark channel prior (DCP) [42] in image defogging, this idea is also used in underwater restoration due to the similarity of underwater and haze scenes to some extent. Based on this idea, Drews et al. proposed underwater DCP (UDCP) [43], which only used blue and green channels. Peng et al. [44] proposed an underwater depth estimation method based on image blurriness and light absorption. Berman et al. [45]considered multiple spectral profiles for different water types. Muniraj et al. [46] used a guided filter and rolling guidance filter to refine the depth map. However, in some cases of severe attenuation, these methods don't yield satisfactory results.

Given the success of DL in the atmosphere environment, it is expected to perform well in underwater depth estimation. However, compared with the atmosphere datasets, underwater datasets containing depth ground truth are difficult to obtain. Until now, there is still not enough dataset to support the training of a highly generalized supervised neural network. Some works [47–49] tried to synthesize fake underwater images from land datasets, such as NYU datasets. For example, Ye et al. [47] proposed an unsupervised style-adaptive depth estimation network that adapts in-air images to the style of the underwater domain. Hambarde et al. [50] directly used synthetic underwater images to train an underwater generative adversarial network (UW-GAN), and so on [51,52]. The limitation of these methods is that the types of synthetic underwater images are far from enough to meet the needs of generalization. Thus, estimating depth by unsupervised or self-supervised learning is essential [48] [53]. In 2019, Gupta et al. [48] proposed an unsupervised monocular depth estimation network UW-Net, which learned a mapping function between unpaired RGB-D ground images and arbitrary underwater images to estimate the desired depth map. Self-supervised depth estimation has achieved remarkable success in the atmosphere environment. However, it is very different between the atmosphere and underwater environment in many aspects. Particularly, problems such as low contrast, complicated illumination conditions, and living organisms, are dominated in the underwater cases. All of these factors make underwater depth estimation a challenging task [54].

Stimulated by the depth estimation methods based on the underwater physical model, this paper creatively guides the network itself to learn the depth-changing trend from attenuation images. Specifically, underwater enhancement is applied to promote performance. At the same time, the geometric constraints between the depth map and optical flow and image gradient are used to refine the depth map. Moreover, the optical flow consistency check is deployed to deal with the occlusion problem caused by moving targets.

Compared with other DL-based depth estimation methods, this work effectively alleviates the problems of insufficient paired RGB-D datasets, lacking accurate calibration, and poor generalization, presenting a great application prospect in underwater depth estimation.

3. Methods

3.1. Main principle

The core underlying idea is considering depth estimation as a view-synthesis process.

Reprojection: Suppose I_s and I_t are two consecutive RGB frames sampled from a raw video. D_t is the estimated dense depth of I_t from the depth network with trainable parameters θ. $T_{t\to s}$ is the relative transformation of I_s and I_t, which is estimated by a pose network with trainable parameters ϕ. K is the camera intrinsic, which is usually known. With the above variables, the view-synthesis process can be described as the following formulation,

$$p_s \sim KT_{t\to s}D(p_t)K^{-1}p_t, \tag{1}$$

where p_t is the arbitrary pixel in I_t, and p_s is the pixel in I_s corresponding to p_t. Based on the above per-pixel correspondence, I_t can be reconstructed as $\hat{I}_t$ from I_s with differentiable bilinear sampling [55].

Reconstruction error: The network (depth network and pose network) learning is based on the above warping process. The objective is to reduce reconstruction error by optimizing θ and ϕ to obtain more accurate results. Usually, the metric to quantify the reconstruction error is the photometric error [56,39].

Regularization: Supervision information solely dependent on reconstruction error is not enough due to one pixel may match many candidates. One commonly used method is smoothness regularization, which is proposed in the previous work [57]. The purpose is to encourage the estimated depth to be locally similar when no significant image gradient exists [36].

3.2. Underwater characteristics

Low Contrast: The fundamental feature of the underwater scene is the low contrast and heavy noise [58]. All the image processing algorithms such as feature matching are suffered from blur. Moreover, the calculated photometric losses of these blurry images are either very small or incorrect. Thus, the network lacks the right direction to optimize. The intuitive idea is to estimate the depth through the enhanced or restored image directly. However, the trained model in this way still can not be generalized to blurred images.

Fortunately, there is a strong correlation between depth-changing and underwater light attenuation, which can help us turn a disadvantage into an advantage. Based on this idea, the potential optical attenuation information in the original images fed into the network should be preserved completely. At the same time, we hope that the network can learn an ability to explore deep abstract features from attenuation images through the guidance of strong loss. Thus, the original blurred images are only used in the forward propagation process, while the loss is calculated on the corresponding pixels in the enhanced images. At the same time, we propose a robust gradient-based loss to refine the texture details and eliminate the effect of attenuated photometric loss. The key idea is to calculate the photometric loss in the gradient domain of target images and reconstructed images. CLAHE [59] is applied as the image enhancement method in this work. The Sobel operator is applied to obtain the gradient map. Some examples of gradient maps and enhanced images are shown in Fig. 1.

Uneven Illumination: Different from the terrestrial environment, the illumination of the underwater environment varies widely [60]. In underwater scenes, due to the strong scattering effect, the illumination varies heavily with the perception angle and different scenes [61]. Such uneven and changeable illumination features will lead to the fatal failure where the background is considered as the foreground mistakenly. Seeking additional robust constraints is therefore necessary for training a better depth estimation model.

The problem can be constrained by imposing the following prior knowledge of underwater scenes: The background should be far away. Based on this prior, we seek a robust constraint by optical flow. Specifically, the disparity values of background regions generated by optical flow are usually small, as shown in Fig. 2. When the depth network incorrectly predicts background regions as small depths, it means that the disparity values of these regions are correspondingly large. At this time, the difference of disparity values generated by optical flow and depth network can provide a penalty term for the depth network. Among them, the difference of disparity values can be represented by pixel-space and camera-space.

Dynamic Objects: Moving objects usually have a large displacement, which violates the static world assumption. For the dynamic objects in the road scene, semantic segmentation is a feasible solution. The types of dynamic objects on the road are relatively simple, such as vehicles and pedestrians. However, abundant types of organisms are moving in the underwater environment [62] [63]. It is very difficult to train a segmentation model for all moving targets [64].

Optical flow, as a motion description of images, has been widely used in high-level tasks and presents great potential in underwater perception. In this paper, the optical flow is used to deal with these ambiguous regions. Specifically, the consistency checks of optical flow and pixel values are used to generate occlusion masks. Then the occlusion regions can be masked to alleviate the influence of moving objects.

3.3. Network architecture

Combined with the above analyses, an underwater self-supervised depth estimation framework is designed as shown in Fig. 3. It consists of three major components, *i.e.* DepthNet, PoseNet, and FlowNet.

DepthNet, which is the encoder-decoder architecture with skip connections, takes a single raw image as input and predicts a depth map. The architecture of DepthNet is shown in Fig. 4(a). Among that, the encoder is based on ResNet18 pre-trained on ImageNet [65], which is lighter and faster. The decoder uses skip connections from the encoder's activation blocks, enabling the network to represent both deep abstract features as well as local information. We applied one convolution layer and sigmoid function to the decoder network, generating disparity predictions at four different scales.

The PoseNet takes consecutive frames as input and predicts the relative transformation between the source image and target image. The architecture of PoseNet is shown in Fig. 4(b). Similarly, the pre-trained ResNet18 is applied to extract deep features. Then, three convolution layers are connected in turn to integrate the pose information. The output of PoseNet is the 6-DOF transformation between the input frames, which is represented as 6 numbers: (T_x, T_y, T_z) for the translation and (R_x, R_y, R_z) for the rotation using the Euler parameterization.

The FlowNet is based on the FlowNet2.0 architecture introduced by [66]. In this work, we use the pre-trained model.

3.4. Optimization objective

Training is driven by self-supervised constraints, including photometric constraint, smoothness constraint, consistency constraint, and gradient constraint. The total loss is the weighted sum of these components.

Photometric constraint $\mathscr{L}_p$: This constraint is the similarity measure. Following [56], the definition is as follows:

$$\mathscr{L}_p(I'_t, \hat{I}'_t) = \frac{\alpha}{2}(1 - SSIM(I'_t, \hat{I}'_t)) + (1-\alpha)\left\|I'_t - \hat{I}'_t\right\|_1 \cdot M \tag{2}$$

I'_t and $\hat{I}'_t$ denote the enhanced target image and reconstructed image, respectively. *SSIM* denotes the structural similarity index and $\|\cdot\|_1$ denotes the L1 loss. The combination of L1 and SSIM measures the appearance similarity between I'_t and $\hat{I}'_t$. The parameter α is taken to be 0.85. To avoid the influence of occlusion regions, the forward–backward consistency checks of optical flow and color images are used to generate the mask M of non-occlusion regions [67], which is described as follows:

$$M = [w^f(x) + w^b(x + w^f(x))] < Th \quad \cap \quad [I_1(x) - I_2(x + w^f(x))] < Th \tag{3}$$

I_1 and I_2 present the images. x presents position on the pixel coordinate. w^f denotes an optical flow that goes from I_1 to I_2, and w^b is the one with the opposite direction. Th is set as 1 in this work.

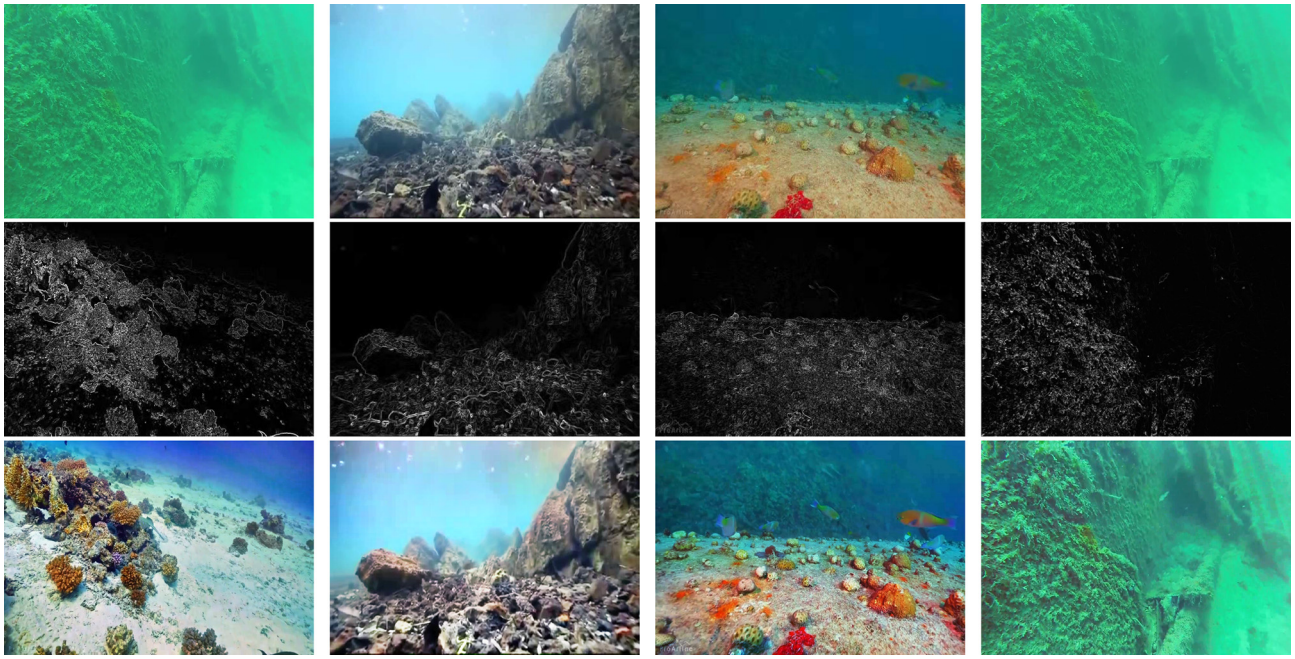


Fig. 1. Several examples of gradient maps and enhanced images. The first row is the original underwater images. The second row is the corresponding gradient map. The third row is the corresponding enhanced images.

Fig. 2. Several examples of optical flow. The first and second rows are the original images. The third row is the corresponding optical flow.

In mask M, the values in the occluded regions are 0 and the values in the non-occluded regions are 1.

Smooth constraint $\mathscr{L}_s$: We follow previous work [56] by applying edge-aware smoothness loss to enforce smoothness in depth. The definition of this constrain is as follows:

$$\mathscr{L}_s = \sum_{p_t} |\bigtriangledown D(p_t)| \cdot \left(e^{-|\bigtriangledown I'_{p_t}|}\right)^T \tag{4}$$

$D(p_t)$ is the estimated depth at pixel p_t. $|\cdot|$ denotes element-wise absolute value. $\bigtriangledown$ is the vector differential operator. T denotes the transpose of image gradient weighting.

Consistency constraint $\mathscr{L}_c$: Following [68], the consistency constrain $\mathscr{L}_c$ is the sum of two components, which are pixel-space consistency $\mathscr{L}_{ci}$ and camera-space consistency $\mathscr{L}_{cc}$.

The definition of $\mathscr{L}_{ci}$ is described as follows:

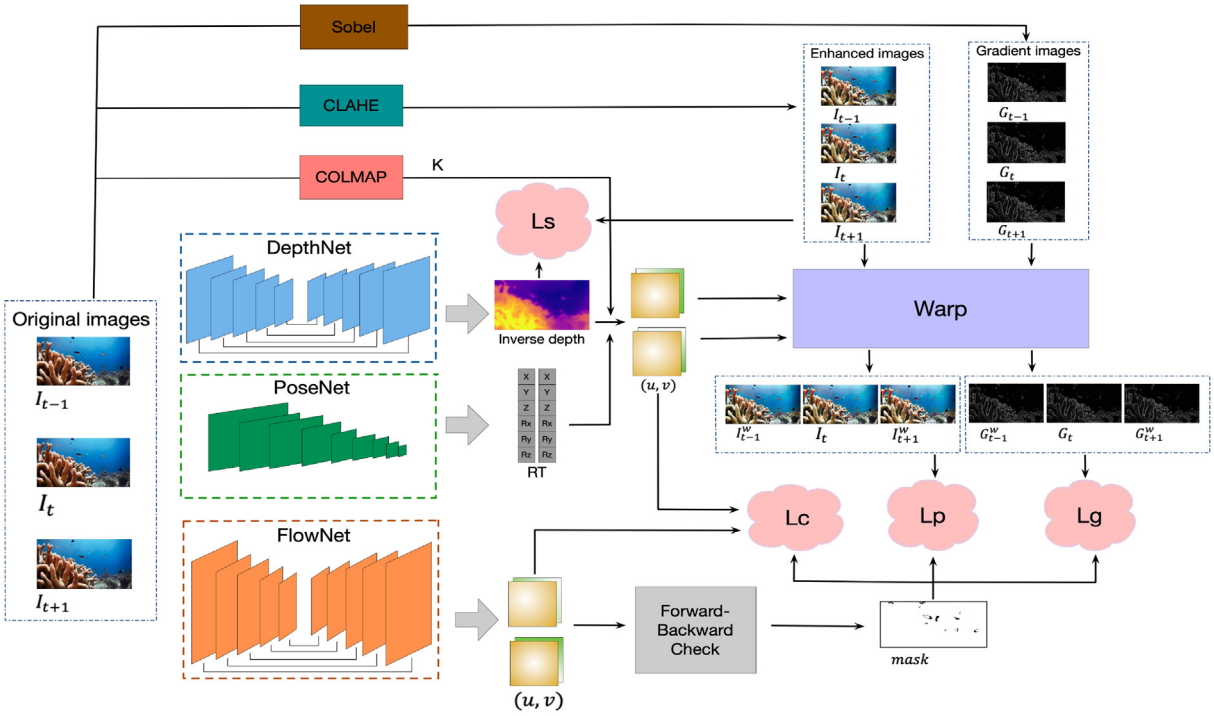


Fig. 3. The framework of proposed underwater self-supervised depth estimation.

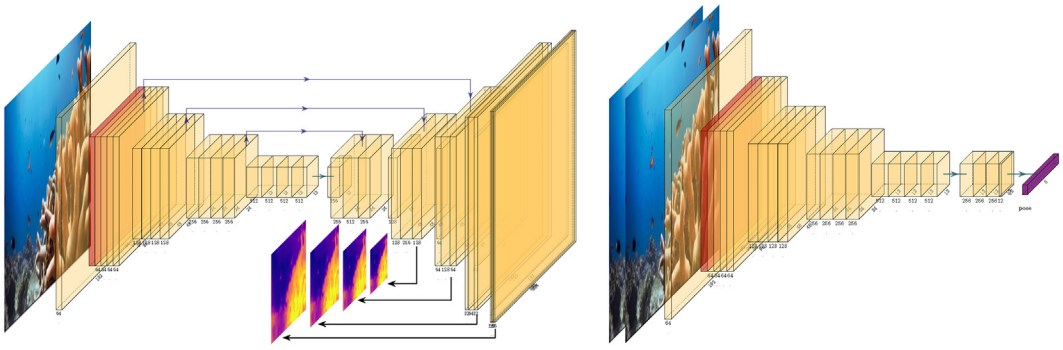


Fig. 4. The architecture of DepthNet and PoseNet.

$$\mathscr{L}_{ci} = \sum_{p_t} \left\| p_s^d - p_s^f \right\| \cdot M \tag{5}$$

p_s^d denotes the 2D-coordinate in the source image corresponding to the pixel p_t in the target image, which is calculated by depth and pose. p_s^f denotes the 2D-coordinate in the source image corresponding to the pixel p_t in the target image, which is calculated by optical flow. The intuitive display of $\mathscr{L}_{ci}$ is shown in Fig. 5(a).

The definition of $\mathscr{L}_{cc}$ is described as follows:

$$\mathscr{L}_{cc} = \sum_{p_t} \left\| P_s^d - P_s^f \right\| \cdot M \tag{6}$$

P_s^d denotes the 3D-coordinate in the source camera corresponding to the pixel p_t in the target image, which is calculated by depth and pose. P_s^f denotes the 3D-coordinate in the source camera corresponding to the pixel p_t in the target image, which is calculated by optical flow. The intuitive display of $\mathscr{L}_{cc}$ is shown in Fig. 5(b).

Gradient constraint $\mathscr{L}_g$: The definition of gradient constraint is consistent with that of photometric constraint, except for the I'_t and $\hat{I}'_t$. This loss is described as follows:

$$\mathscr{L}_g(g_t, \hat{g}_t) = \frac{\alpha}{2}(1 - SSIM(g_t, \hat{g}_t)) + (1-\alpha)\|g_t - \hat{g}_t\|_1 \cdot M \tag{7}$$

g_t and $\hat{g}_t$ denote the first-order gradient of the target image I_t and reconstructed image $\hat{I}_t$, respectively.

Final Loss:

In summary, the total loss used in this work is as follows:

$$\mathscr{L} = \mathscr{L}_p + \lambda_s \mathscr{L}_s + \lambda_c \mathscr{L}_c + \lambda_g \mathscr{L}_g, \tag{8}$$

where, λ_s, λ_c, and λ_g denote loss weight, respectively. In this paper, we set $\lambda_s = 1e-4, \lambda_c = 8e-3, \lambda_g = 1$, empirically.

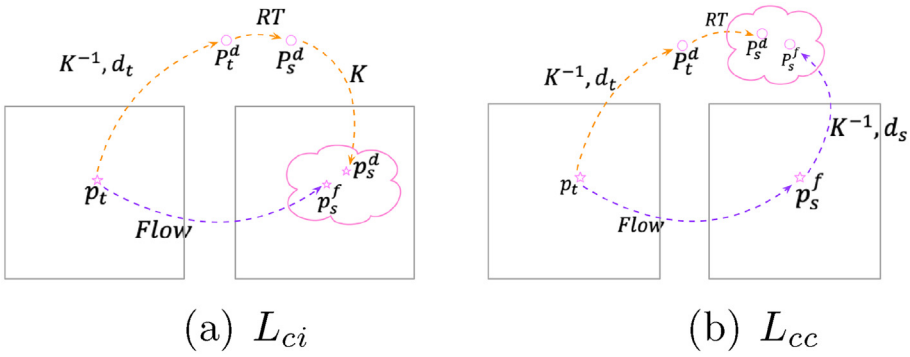


Fig. 5. The intuitive representation of consistency between depth and optical flow.

4. Experiments

4.1. Datasets

Most existing self-supervised depth estimation models are trained on the standard air medium datasets, such as KITTI [21,69], CityScape [70], *etc*. However, underwater images often undergo distortions from scattering, absorption, color loss, diffraction, and varying attenuation depending on the light frequency. The datasets distributions of air and underwater images are usually quite different. Thus, the models trained in air medium datasets are not able to generalize to underwater images. This prompts us to construct new training datasets suitable for training underwater depth estimation. We believe that this is of great significance to promote the development of underwater depth estimation.

The datasets come from underwater videos on the internet. There are 51 videos of the underwater environments, including 12438 frames. The datasets provide a relatively comprehensive and vivid display of the underwater world, including many objects such as swimming fish, swinging aquatic plants, corals, and pebbles. Some examples of datasets are shown in Fig. 1. The underwater characteristics are clearly illustrated from the datasets. First of all, the areas of weak texture regions in underwater datasets are larger than those in atmosphere images, especially reflected in the background. Then, sizeable different types of dynamic objects exist underwater, such as fish, aquatic plants, and many other organisms. Furthermore, under the combined action of refraction and scattering by water, the collected underwater datasets have relatively low contrast.

Notably, in the framework of self-supervised depth estimation, the intrinsic K is a fundamental parameter, which connects the relationship between the pixel space and camera space of the same scene. However, the collected underwater videos from the internet lack information about the camera. The COLMAP [71], a state-of-the-art multi-view stereo method, is applied to calculate the intrinsic parameters.

4.2. Experimental setup and metrics

Implement Details: Our network is implemented using the PyTorch library [72] and trained on a single GeForce 1080. We jointly optimize both DepthNet and PoseNet with the Adam Optimizer and a learning rate of $1e-4$. Image resolution is set to 384×384 pixels.

Metrics: We adopt nine evaluation metrics to quantitatively evaluate. They are Pearson correlation coefficient (ρ), log scale-invariant mean squared error (*SI-MSE*), absolute relative error(*Abs Rel*), square relative error(*Sq Rel*), root mean square error(*RMSE*), root mean square logarithmic error(*RMSE log*), accuracy with threshold ($\delta < thr$). Among them, ρ and *SI-MSE* are common metrics for underwater depth estimation [48]. The residual metrics are often used to evaluate the depth estimation of the atmospheric environment [56]. Smaller values of *SI-MSE*, *Abs Rel*, *Sq Rel*, *RMSE*, and *RMSE log* indicate improved performance. Larger values of ρ and $\delta < thr$ indicate improved performance.

4.3. Comparison with SOTA methods

4.3.1. Compared methods

We compare our work with other state-of-the-art depth estimation algorithms from two main categories.

- In the physical model-based image restoration, the depth map is an intermediate link. Several representative related works are as follows: MIP [40], UDCP [43], HL [45], Peng [44], Peng [73].
- DL-based depth estimation methods: UW-Net [48], Mono-Depth2 [56], Walvelet-monodepth [74], HR-depth [26], and ManyDepth [39]. Among them, the UW-Net [48] is an unsupervised method, which was proposed for underwater depth estimation. While other methods are self-supervised methods. The original purpose of them is to deal with atmosphere scenes. All the competitors are trained on the same underwater training datasets for a fair comparison.

4.3.2. Qualitative analysis

The results of all the competitors are shown in Fig. 6, while (b-f) are physical model-based methods and (g-l) are DL-based methods. The results are colored to improve the visualization, where, warm colors represent closer points and cool colors represent farther points.

As shown in Fig. 6(b-f), the results of physical model-based methods present more detailed information. However, the trends of depth-changing are incorrectly estimated in some scenes, such as the last column in Fig. 6(b-f), where the shipwreck is incorrectly labeled as the far object in most methods. Among all the tested physical model-based methods, the results of MIP are characterized by its rich details, as shown in Fig. 6(b). However, even MIP fails to estimate the correct depth-changing trends in all types of data. A total of 30% of the data has the above issue.

Based on the DCP, which was proposed to deal with haze, UDCP [43] only applied the green and blue channels of underwater images due to the difficulty of modeling the behavior of the red channel. The colored results of UDCP are shown in Fig. 6(c). It is easy to see that some data still have ambiguous problems in foreground and background regions. As shown in Fig. 6(d), Peng [44] can not generate correct depth-changing trends in most data. Compared with [44], Peng [73] has significant improvement. However,

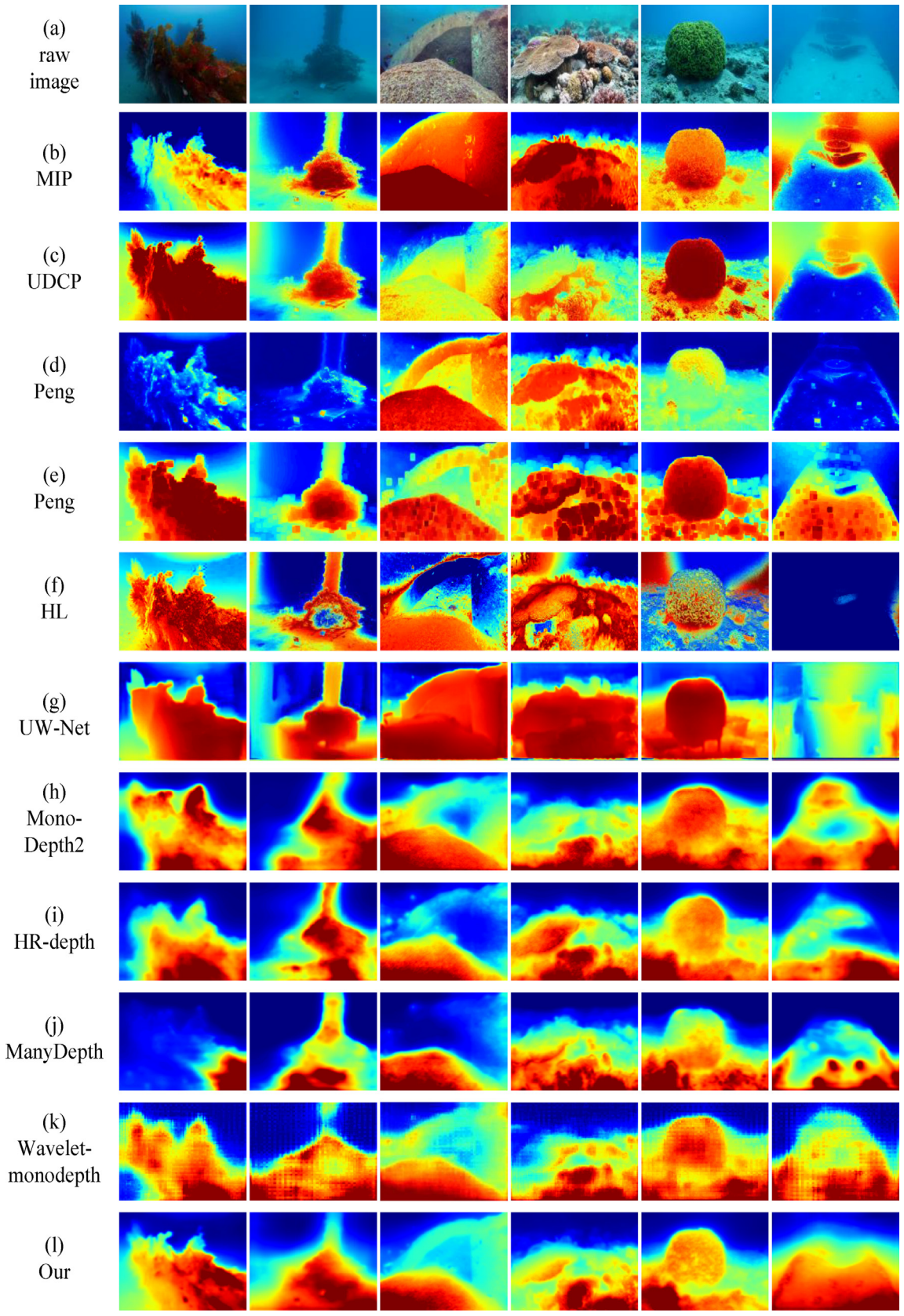


Fig. 6. Qualitative comparison results of our method and other state-of-the-art depth estimation methods.

Table 1
Quantitative comparison of our method and the state-of-the-art methods in the SQUID dataset.

Methods		$\rho\uparrow$	*SI-MSE* ↓	*Abs Rel* ↓	*Sq Rel* ↓	*RMSE* ↓	*RMSE log* ↓	$\delta<1.25\uparrow$	$\delta<1.25^2\uparrow$	$\delta<1.25^3\uparrow$
Physical model-based										
UDCP [43]		0.027	0.567	9.153	5.683	0.466	1.854	0.092	0.178	0.257
MIP [40]		0.380	0.463	8.193	4.946	0.434	1.744	0.091	0.173	0.257
Peng [44]		0.601	0.428	11.614	8.417	0.631	2.081	0.022	0.065	0.121
Peng [73]		0.574	0.458	9.467	5.692	0.481	1.908	0.061	0.121	0.189
HL [45]		0.252	0.522	10.231	7.503	0.522	1.853	0.084	0.171	0.256
DL-based										
UW-Net		0.538	0.378	5.454	2.713	0.294	1.414	**0.139**	**0.264**	**0.382**
MonoDepth2 [56]	aug	0.352	0.422	5.404	2.258	0.368	1.576	0.088	0.170	0.260
	no aug	0.764	0.245	4.275	1.458	0.324	1.402	0.082	0.179	0.280
HR-depth [26]	aug	0.558	0.301	4.613	1.695	0.332	1.470	0.112	0.212	0.298
	no aug	0.790	0.229	4.239	1.446	0.310	1.381	0.092	0.186	0.287
ManyDepth [39]	aug	0.641	0.252	4.792	1.811	0.361	1.504	0.076	0.153	0.238
	no aug	0.740	0.229	4.892	2.064	0.390	1.489	0.063	0.142	0.244
Walvelet- monodepth [74]	aug	0.415	0.327	5.856	2.811	0.427	1.626	0.068	0.143	0.226
	no aug	0.791	0.231	3.758	1.093	0.274	1.316	0.106	0.211	0.325
Our	no aug	**0.812**	**0.206**	**3.396**	**0.892**	**0.262**	**1.271**	0.120	0.221	0.330

Table 2
The results of ablation study.

	Enh BP	**Enh FP**	**Gradient constraint**	**Consistency constraint**	$\rho\uparrow$	*SI-MSE* ↓	*Abs Rel* ↓	*Sq Rel* ↓	*RMSE* ↓	*RMSE log* ↓	$\delta<1.25\uparrow$	$\delta<1.25^2\uparrow$	$\delta<1.25^3\uparrow$
1	×	×	×	×	0.764	0.245	4.275	1.458	0.324	1.402	0.082	0.179	0.280
2	✓	✓	×	×	0.740	0.251	4.549	1.817	0.350	1.418	0.084	0.177	0.277
3	✓	×	×	×	0.777	0.236	3.987	1.357	0.307	1.343	0.107	0.216	0.327
4	✓	×	×	✓	0.822	0.199	3.046	0.739	0.238	1.201	0.125	0.238	0.366
5	✓	×	✓	✓	0.812	0.206	3.396	0.892	0.262	1.271	0.120	0.221	0.330

[1] **Enh BP** indicates underwater image enhancement is applied in backward propagation.
[2] **Enh FP** indicates underwater image enhancement is applied in forward propagation.

some data still have the problem of the wrong estimation, as shown in the second column in Fig. 6(e). The results of HL are not ideal, which may be because the library of water types used in this paper can not be functional for different water quality.

Regarding to the DL-based methods, a common problem is that detailed information is missed in most scenes. All the comparative methods show kinds of defects. Specifically, UW-Net is an unsupervised underwater monocular depth method. It presents certain robust facing underwater scenes. But there are also some problems worth discussing. In some scenarios, the trend of distance variation is not correctly estimated, such as the last column in Fig. 6(g). As the classical self-supervised monocular depth estimation pipeline, the MonoDepth2 presents certain robustness for the underwater environment. The results reflect the general depth-changing trends but lack detailed information. The results of ManyDepth and HR-depth are similar to MonoDepth2. Wavelet-monodepth exploits wavelet decomposition, which is integrated with a fully differentiable encoder-decoder architecture, to achieve optimal efficiency. As shown in Fig. 6(k), this makes the depth-changing trends clearer than MonoDepth2.

The colored results of our method are shown in Fig. 6(l). Compared with the above methods, the depth variation trends obtained from our results are more accurate than all the other compared methods. What's more, our method also performs well in some intractable scenarios, such as the last column in Fig. 6(l). This indicates that our method presents a more robust performance in the face of blurred environment. Further, the detail retention of our method is better than all the comparative DL-based methods. This is because our network effectively combines image enhancement and gradient to refine the texture information.

4.3.3. Quantitative analysis

We evaluate all the methods on SQUID datasets [45], which contain four scenes of stereo underwater images with corresponding depth ground truth. Following UW-Net [48], we used three scenes, which are Katzaa, Nachsholim, and Satil, (72 images in total) to evaluate the results. Considering that the depth maps generated by the stereo camera are not complete, we only use the effective pixels for evaluation. The comparison results are listed in Table 1.

In general, DL-based depth estimation methods generated better performance in each evaluation metric. This is because the depth tendencies of most of them are more accurate than physical model-based methods. However, the physical model-based methods can retain more details in the results which is not been fully represented in depth estimation metrics. Generally, data augmentation is used to improve the performance in self-supervised depth estimation of atmosphere. In this paper, we also explore the impact of data augmentation on underwater depth estimation through comparative experiments. The results are also listed in Table 1. Experimental results show that data augmentation has no positive effect on self-supervised underwater depth estimation. This may be because the difference in data distribution between augmented datasets and original underwater datasets is too large. Thus, we discarded this operation in our work. Compared with other methods, both the proposed method and UW-Net achieve considerable results. UW-Net obtains the best performance in the last three evaluation metrics. Our method significantly outperforms other competitors in other evaluation metrics. This is also consistent with the subjective results.

4.4. Ablation study

Here, based on MonoDepth2, we gradually add the different modules to verify the effect of each part on underwater depth estimation. The quantitative comparison results are listed in Table 2. The qualitative comparison results are shown in Fig. 7. Among that, the first row in Table 2 and Fig. 7(b) are the results of MonoDepth2 in underwater images.

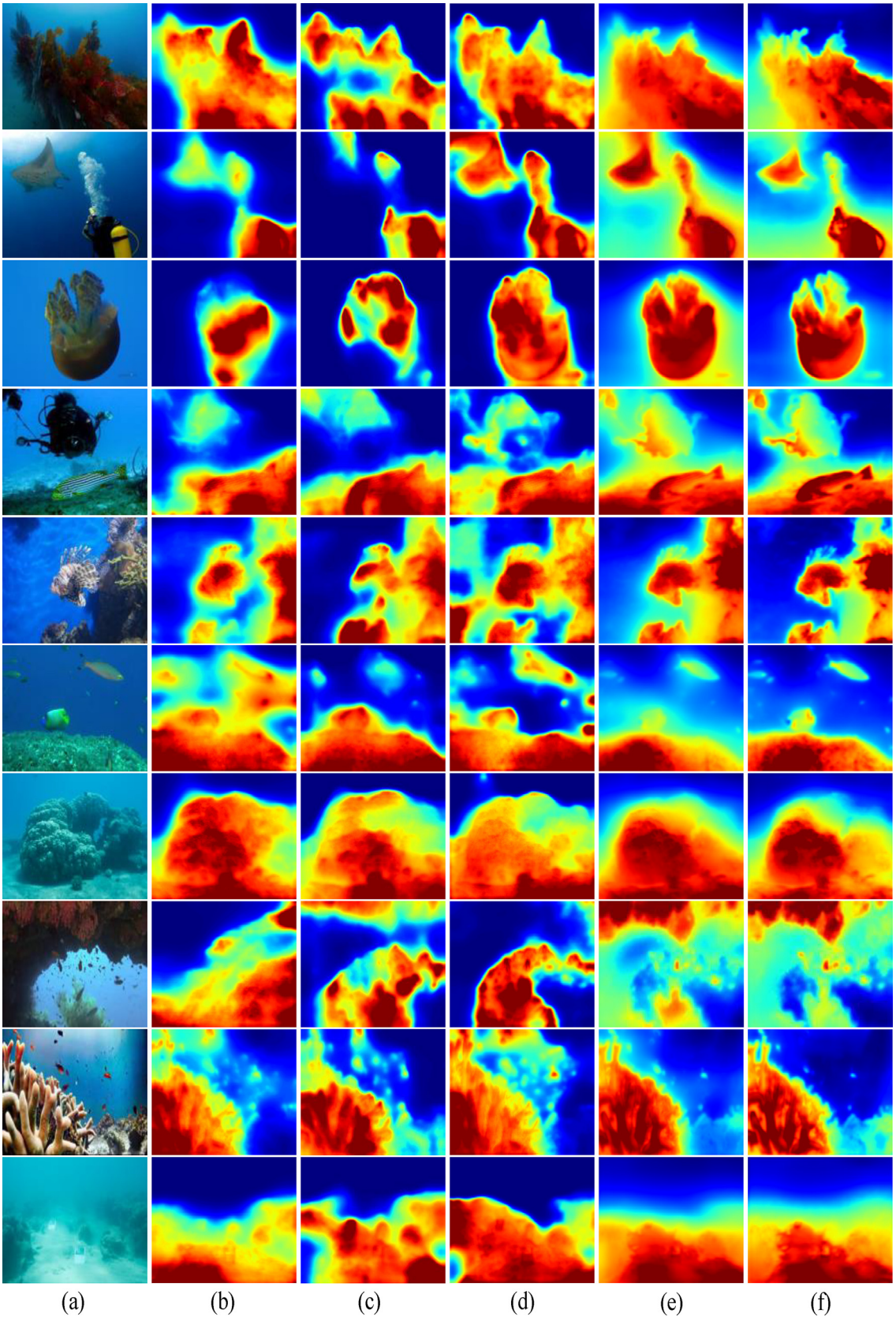


Fig. 7. Qualitative comparative results of ablation experiments.

The depth-changing trend and underwater light attenuation are closely correlated. We think that light attenuation presented in the original underwater images would be helpful for depth map inference. Thus, we only use underwater image enhancement in backward propagation. By this way, the potential depth information in the original underwater images is not destroyed, and the gener-

Table 3
Quantitative results for different backbone networks.

Methods	$\rho\uparrow$	*SI-MSE* ↓	*Abs Rel* ↓	*Sq Rel* ↓	*RMSE* ↓	*RMSE log* ↓	$\delta < 1.25\uparrow$	$\delta < 1.25^2\uparrow$	$\delta < 1.25^3\uparrow$
VGG19	0.469	0.340	4.309	1.617	0.309	1.382	**0.125**	**0.244**	**0.355**
DenseNet121	0.793	**0.194**	3.748	1.136	0.305	1.334	0.084	0.182	0.289
ResNet18	**0.812**	0.206	**3.396**	**0.892**	**0.262**	**1.271**	0.120	0.221	0.330

alization of the network is improved. We made comparative experiments to validate this idea. Firstly, we use underwater image enhancement in both forward and backward propagation. Then, we only use image enhancement in backward propagation. The quantitative comparison results are listed in the second and third row of Table 2, respectively. Compared with the results of Mono-Depth2, which is listed in the first row, we can see a drop presented in most metrics obtained from the first way. When only applying image enhancement in backward propagation, the depth estimation results have improved in all metrics. The qualitative results are shown in Fig. 7(c) and (d), respectively. We can observe that the model trained from the first way does not learn the ability to dig effective information from the original images, including distance and texture information. However, through the guidance of image enhancement in backward propagation, the network learns the ability to infer more accurate depth maps from attenuated images.

Although the underwater image enhancement technology can guide the network, it will also magnify some inherent problems in underwater images, such as noise and uneven illumination, as shown in the fifth row of Fig. 7(d). Therefore, we further introduce the consistency between optical flow and depth to constraint the network. The experimental results are shown in Fig. 7(e). It is obvious that some erroneously estimated areas are effectively corrected. At the same time, in some scenarios where it is difficult for the network to distinguish between foreground and background, the consistency constraint plays a prominent role, such as the third row to the last in Fig. 7(e). The quantitative results are listed in the fourth row of Table 2. We can see a sudden rise in all metrics, which also indicates the effectiveness of consistency constraint between optical flow and depth.

The combination of the above modules has achieved considerable results in underwater images. However, texture information is not well preserved. Thus, we introduce the gradient loss on the basis of the above experimental design as our Full Method. The quantitative results are listed in the last row of Table 2 and the quantitative results are shown in Fig. 7(f). From the perspective of evaluation metrics, the score of Full Method decreases. However, in the pictures in Fig. 7(f), we can see that estimated depth maps obtain shaper edges and richer details.

In summary, underwater image enhancement, consistency constraint between depth and optical flow, and image gradient play the important role in improving the results of underwater depth estimation. They improve the underwater depth estimation results from different perspectives. Image enhancement helps the network explore distance variations from attenuated images. The consistency between optical flow and depth can help the network alleviate some inherent problems in underwater scenes, such as uneven illumination, noise, and so on. The image gradient can help the network refine texture details.

4.5. Choice of backbone network

For validating the results of different backbones, we replace the ResNet18 encoder with different backbone networks, which are VGG19 [75] and DenseNet121 [76]. The quantitative results of different backbones are shown in Table 3. As we can see in this table, ResNet18 performs best on most metrics. VGG19 outperforms ResNet18 in the last three metrics. DenseNet121 has the best result in *SI-MSE* metric. In this paper, we choose ResNet18 as the backbone network.

5. Conclusion

In this paper, a novel self-supervised monocular underwater depth estimation framework is proposed based on comprehensive analyses of the underwater characteristics. This pipeline makes full use of the potential clues in attenuation images to infer depth-changing trends with the guidance of multiple loss functions. In addition, the image gradient and consistency between depth and optical flow are used to further refine the depth map.

Depth estimation is vital in autonomous underwater exploration. We believe this work provides a practical solution, which can serve in most underwater robots without adding any extra apparatus. However, our framework still presents some limitations. Firstly, the details of the depth map are not fully translated from the original images. Then, driven by the goal of underwater *real-time* and *in situ* perception, our model needs to be further simplified in future work.

CRediT authorship contribution statement

Xuewen Yang: Investigation, Validation, Writing – original draft. **Xing Zhang:** Formal analysis, Data curation, Writing – original draft. **Nan Wang:** Conceptualization, Supervision, Visualization, Writing – review & editing, Methodology, Funding acquisition. **Guoling Xin:** Formal analysis, Data curation. **Wenjie Hu:** Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61703381, U2006228).

References

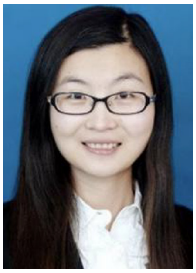
[1] G. Picardi, M. Chellapurath, S. Iacoponi, S. Stefanni, C. Laschi, M. Calisti, Bioinspired underwater legged robot for seabed exploration with low environmental disturbance, Sci. Robot. 5 (42).
[2] Z. Gong, X. Fang, X. Chen, J. Cheng, Z. Xie, J. Liu, B. Chen, H. Yang, S. Kong, Y. Hao, et al., A soft manipulator for efficient delicate grasping in shallow water: Modeling, control, and real-world experiments, Int. J. Robot. Res. 40 (1) (2021) 449–469.
[3] M. Johnson-Roberson, M. Bryson, A. Friedman, O. Pizarro, G. Troni, P. Ozog, J.C. Henderson, High-resolution underwater robotic vision-based mapping and three-dimensional reconstruction for archaeology, J. Field Robot. 34 (4) (2017) 625–643.
[4] J. Sverdrup-Thygeson, E. Kelasidi, K.Y. Pettersen, J.T. Gravdahl, The underwater swimming manipulator–a bioinspired solution for subsea operations, IEEE J. Oceanic Eng. 43 (2) (2017) 402–417.
[5] W.-S. Choi, D.R. Olson, D. Davis, M. Zhang, A. Racson, B. Bingham, M. McCarrin, C. Vogt, J. Herman, Physics-based modelling and simulation of multibeam

echosounder perception for autonomous underwater manipulation, Frontiers in Robotics and AI 8.
[6] K. Pathak, A. Birk, N. Vaskevicius, Plane-based registration of sonar data for underwater 3D mapping, IROS (2010) 4880–4885.
[7] Y. Guo, 3D underwater topography rebuilding based on single beam sonar, ICSPCC (2013) 1–5.
[8] E. Coiras, Y. Petillot, D.M. Lane, Multiresolution 3-D reconstruction from side-scan sonar images, IEEE TIP 16 (2) (2007) 382–390.
[9] L.K. Rumbaugh, E.M. Bollt, W.D. Jemison, Y. Li, A 532 nm chaotic lidar transmitter for high resolution underwater ranging and imaging, OCEANS (2013) 1–6.
[10] G. Kim, Y. Lim, J. Park, W. Kim, D. Lee, H. Cho, C. Park, S. Kang, K. Kim, S. Park, et al., Single-energy material decomposition in radiography using a three-dimensional laser scanner, J. Korean Phys. Soc. 75 (2) (2019) 153–159.
[11] P. Mariani, I. Quincoces, K.H. Haugholt, Y. Chardard, A.W. Visser, C. Yates, G. Piccinno, G. Reali, P. Risholm, J.T. Thielemann, Range-gated imaging system for underwater monitoring in ocean environment, Sustainability 11 (1) (2019) 162.
[12] S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud, CVPR, 2019, pp. 770–779.
[13] Y. Wu, Y. Zhou, S. Chen, Y. Ma, Q. Li, Defect inspection for underwater structures based on line-structured light and binocular vision, Appl. Opt. 60 (25) (2021) 7754–7764.
[14] L.R. Ramírez-Hernández, J.C. Rodríguez-Quinoñez, M.J. Castro-Toscano, D. Hernández-Balbuena, W. Flores-Fuentes, R. Rascón-Carmona, L. Lindner, O. Sergiyenko, Improve three-dimensional point localization accuracy in stereo vision systems using a novel camera calibration method, IJARS 17 (1) (2020), 1729881419896717.
[15] M. Mustonen, A. Klauson, M. Andersson, D. Clorennec, T. Folegot, R. Koza, J. Pajala, L. Persson, J. Tegowski, J. Tougaard, et al., Spatial and temporal variability of ambient underwater sound in the baltic sea, Scientific Rep. 9 (1) (2019) 1–13.
[16] N. Wang, H. Zheng, B. Zheng, Underwater image restoration via maximum attenuation identification, IEEE Access 5 (2017) 18941–18952.
[17] D. Akkaynak, T. Treibitz, Sea-thru: A method for removing water from underwater images, CVPR (2019) 1682–1691.
[18] W. Song, Y. Wang, D. Huang, D. Tjondronegoro, A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration, in: PCM, Springer, 2018, pp. 678–688.
[19] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, arXiv preprint arXiv:1406.2283.
[20] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, ECCV (2012) 746–760.
[21] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237.
[22] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, in: 3DV, 2016, pp. 239–248.
[23] W. Zhuo, M. Salzmann, X. He, M. Liu, Indoor scene structure analysis for single image depth estimation, CVPR (2015) 614–622.
[24] D. Xu, E. Ricci, W. Ouyang, X. Wang, N. Sebe, Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation, CVPR (2017) 5354–5362.
[25] J. Jiao, Y. Cao, Y. Song, R. Lau, Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss, in: ECCV, 2018, pp. 53–69.
[26] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, Y. Yuan, HR-depth: High resolution self-supervised monocular depth estimation, CoRR.
[27] J. Watson, M. Firman, G.J. Brostow, D. Turmukhambetov, Self-supervised monocular depth hints, CVPR (2019) 2162–2171.
[28] R. Garg, V.K. Bg, G. Carneiro, I. Reid, Unsupervised CNN for single view depth estimation: Geometry to the rescue, ECCV (2016) 740–756.
[29] T. Zhou, M. Brown, N. Snavely, D.G. Lowe, Unsupervised learning of depth and ego-motion from video, CVPR (2017) 1851–1858.
[30] S. Zhu, G. Brazil, X. Liu, The edge of depth: Explicit constraints between segmentation and depth, CVPR, 2020, pp. 13116–13125.
[31] Y. Lu, M. Sarkis, G. Lu, Multi-task learning for single image depth estimation and segmentation based on unsupervised network, ICRA (2020) 10788–10794.
[32] J. Li, J. Zhao, S. Song, T. Feng, Unsupervised joint learning of depth, optical flow, ego-motion from video, arXiv preprint arXiv:2105.14520.
[33] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, M.J. Black, Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation, CVPR (2019) 12240–12249.
[34] Q. Dai, V. Patil, S. Hecker, D. Dai, L. Van Gool, K. Schindler, Self-supervised object motion and depth estimation from video, CVPR (2020) 1004–1005.
[35] Z. Yin, J. Shi, GeoNet: Unsupervised learning of dense depth, optical flow and camera pose, CVPR (2018) 1983–1992.
[36] Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, LEGO: Learning edge with geometry all at once by watching videos, CVPR (2018) 225–234.
[37] L. Andraghetti, P. Myriokefalitakis, P.L. Dovesi, B. Luque, M. Poggi, A. Pieropan, S. Mattoccia, Enhancing self-supervised monocular depth estimation with traditional visual odometry, in: 3DV, 2019, pp. 424–433.
[38] R. Wang, S.M. Pizer, J.-M. Frahm, Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth, CVPR (2019) 5555–5564.
[39] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, M. Firman, The Temporal Opportunist: Self-supervised multi-frame monocular depth, CVPR (2021) 1164–1174.
[40] N. Carlevaris-Bianco, A. Mohan, R.M. Eustice, Initial results in underwater single image dehazing, OCEANS (2010) 1–8.
[41] Y.-T. Peng, X. Zhao, P.C. Cosman, Single underwater image enhancement using depth estimation based on blurriness, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 4952–4956.
[42] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE TPAMI 33 (12) (2010) 2341–2353.
[43] P. Drews, E. Nascimento, F. Moraes, S. Botelho, M. Campos, Transmission estimation in underwater single images, CVPR (2013) 825–830.
[44] Y.-T. Peng, P.C. Cosman, Underwater image restoration based on image blurriness and light absorption, IEEE Trans. Image Process. 26 (4) (2017) 1579–1594.
[45] D. Berman, D. Levy, S. Avidan, T. Treibitz, Underwater single image color restoration using haze-lines and a new quantitative dataset, IEEE TPAMI.
[46] M. Muniraj, V. Dhandapani, Underwater image enhancement by combining color constancy and dehazing based on depth estimation, Neurocomputing 460 (2021) 211–230.
[47] X. Ye, Z. Li, B. Sun, Z. Wang, R. Xu, H. Li, X. Fan, Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks, IEEE TCSVT 30 (11) (2019) 3995–4008.
[48] H. Gupta, K. Mitra, Unsupervised single image underwater depth estimation, ICIP (2019) 624–628.
[49] E.S. Vaz, E.F. de Toledo, P.L. Drews, Underwater depth estimation based on water classification using monocular image, in: 2020 Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics (SBR) and 2020 Workshop on Robotics in Education (WRE), IEEE, 2020, pp. 1–6.
[50] P. Hambarde, S. Murala, A. Dhall, Uw-gan: Single-image depth estimation and image enhancement for underwater images, IEEE Trans. Instrum. Meas. 70 (2021) 1–12.
[51] Q. Zhao, Z. Zheng, H. Zeng, Z. Yu, H. Zheng, B. Zheng, The synthesis of unpaired underwater images for monocular underwater depth prediction, Frontiers in Marine, Science (2021) 1305.
[52] J. Cui, L. Jin, H. Kuang, Q. Xu, S. Schwertfeger, Underwater depth estimation for spherical images, J. Robot. (2021).
[53] K.A. Skinner, J. Zhang, E.A. Olson, M. Johnson-Roberson, Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 7947–7954.
[54] X. Li, H. Hu, L. Zhao, H. Wang, Y. Yu, L. Wu, T. Liu, Polarimetric image recovery method combining histogram stretching for underwater imaging, Scientific Rep. 8 (1) (2018) 1–10.
[55] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Advances in neural information processing systems 28 (2015) 2017–2025.
[56] C. Godard, O. Mac Aodha, M. Firman, G.J. Brostow, Digging into self-supervised monocular depth estimation, in: ICCV, 2019, pp. 3828–3838.
[57] C. Godard, O. Mac Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, CVPR (2017) 270–279.
[58] A. Dudhane, P. Hambarde, P. Patil, S. Murala, Deep underwater image restoration and beyond, IEEE Signal Process. Lett. 27 (2020) 675–679.
[59] J. Ma, X. Fan, S.X. Yang, X. Zhang, X. Zhu, Contrast limited adaptive histogram equalization-based fusion in YIQ and HSI color spaces for underwater image enhancement, IJPRAI 32 (07) (2018) 1854018.
[60] X. Cao, S. Rong, Y. Liu, T. Li, Q. Wang, B. He, Nuicnet: Non-uniform illumination correction for underwater image using fully convolutional network, IEEE Access 8 (2020) 109989–110002.
[61] M. Mathur, N. Goel, Enhancement of nonuniformly illuminated underwater images, Int. J. Pattern Recognit Artif Intell. 35 (03) (2021) 2154008.
[62] M. Mathur, D. Vasudev, S. Sahoo, D. Jain, N. Goel, Crosspooled fishnet: transfer learning based fish species classification model, Multimedia Tools Appl. 79 (41) (2020) 31625–31643.
[63] M. Mathur, N. Goel, Fishresnet: Automatic fish classification approach in underwater scenario, SN Comput. Sci. 2(4).
[64] P.W. Patil, O. Thawakar, A. Dudhane, S. Murala, Motion saliency based generative adversarial network for underwater moving object segmentation, in: 2019 IEEE international conference on image processing (ICIP), IEEE, 2019, pp. 1565–1569.
[65] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, IJCV 115 (3) (2015) 211–252.
[66] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, CVPR (2017) 2462–2470.
[67] Y. Zou, Z. Luo, J.-B. Huang, Df-net: Unsupervised joint learning of depth and flow using cross-task consistency, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 36–53.
[68] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, J. Kopf, Consistent video depth estimation, ACM TOG 39(4) (2020) 71–1.
[69] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, CVPR (2012) 3354–3361.
[70] E. O'Neill, V. Kostakos, T. Kindberg, A. Penn, D.S. Fraser, T. Jones, et al., Instrumenting the city: Developing methods for observing and understanding the digital cityscape, UbiComp, Springer (2006) 315–332.

[71] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: CVPR, 2016, pp. 4104–4113.
[72] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: NIPS, 2017.
[73] Y.-T. Peng, K. Cao, P.C. Cosman, Generalization of the dark channel prior for single image restoration, IEEE Trans. Image Process. 27 (6) (2018) 2856–2868.
[74] M. Ramamonjisoa, M. Firman, J. Watson, V. Lepetit, D. Turmukhambetov, Single image depth prediction with wavelet decomposition, CVPR, 2021, pp. 11089–11098.
[75] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
[76] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

Xuewen Yang received the B.S. degree in electronic information engineering from Weifang University, Weifang, China, in 2018 and the M.S. degree in electronic and communication engineering from Ocean University of China, Qingdao, China, in 2021. Her research interest is underwater image processing.

Xing Zhang received the B.S. degree in information engineering from Langfang Normal University, Langfang, China, in 2019, She is currently pursuing the master's degree in electronic and communication engineering from Ocean University of China, Qingdao, China. Her research interests include deep learning and underwater image processing.



Nan Wang received the B.S. degree in measurement and control technology and instrument, the M.S. degree in instrument science and technology, and the Ph.D. degree in instrument science and technology, all from Southeast University, Nanjing, China, in 2009, 2012, and 2015, respectively. In 2016, she joined the Department of Electronic Engineering, Ocean University of China, Qingdao, China, where she is currently an Associate Professor. Her research interests include underwater image processing, logical stochastic resonance, and robotics.

Guoling Xin received the B.S. degree in communication engineering from the Shandong University of Technology, Zibo, China, in 2020. He is currently pursuing the master's degree in electronic and communication engineering from Ocean University of China, Qingdao, China. Her research interests include deep learning and underwater image processing.

Wenjie Hu received the B.S. degree in electronic information engineering from the Qingdao Agricultural University, Qingdao, China, in 2019. He is currently pursuing the master's degree in electronic and communication engineering from Ocean University of China, Qingdao, China. His research interests include deep learning and underwater image processing.