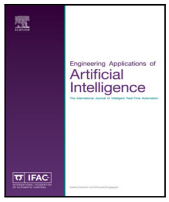
Engineering Applications of Artificial Intelligence 120 (2023) 105846

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Underwater self-supervised monocular depth estimation and its application in image enhancement

Junting Wang, Xiufen Ye *, Yusong Liu, Xinkui Mei, Jun Hou

College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China

ARTICLE INFO

Keywords:
Deep learning
Underwater image enhancement
Underwater depth estimation
Monocular depth estimation

ABSTRACT

Accurate depth estimation is necessary for imaging model-based underwater image enhancement. However, limited by the length of baseline, traditional binocular-based methods are not easy to deploy in underwater scenarios, while monocular methods are more appropriate due to their availability and cost. Therefore, we propose a self-supervised model (called UWdepth) to estimate the depth of underwater scenes with monocular sequences. Addressing the challenges posed by the scale of data and the complexity of motion in underwater scenes, we design an iterative pose network and further introduce a depth consistency loss to achieve more accurate inter-frame motion prediction and depth estimation. The predicted depth can be used to precisely enhance the image based on the underwater imaging model. Experiments show that UWdepth outperforms existing depth estimation models in terms of multiple evaluation indexes. Furthermore, by applying UWdepth, we can enhance underwater images with Akkaynak–Treibitz imaging model, achieving better indexes and visual perception quality than existing traditional and deep-learning based underwater image enhancement algorithms.

1. Introduction

Underwater optical images play an irreplaceable role in underwater visual tasks such as underwater resource exploitation, seabed salvage and environmental monitoring (Raveendran et al., 2021). However, underwater optical images tend to be blue–green in hue and suffer from blurring and confusion due to light absorption and scattering (Li et al., 2022). This poses greater difficulties for the associated underwater high-level visual tasks. Therefore, it is of great significance to enhance the underwater images to improve the quality of visual perception.

It is well known that underwater image enhancement expects to obtain recovered and color corrected images from degraded inputs. Since the ground truth(GT) of underwater image enhancement result is almost impossible to obtain, underwater image enhancement is a challenging task with uncertainty Jian et al. (2021). Researchers have tried to study the imaging model from underwater images, which can simulate precise underwater imaging processes to a certain extent Akkaynak and Treibitz (2018). However, the application of these models is limited by accurate underwater depth information.

In the atmospheric scenes, binocular camera is a common choice for obtaining absolute depth data from images. But the baseline length of the binocular cameras will restrict their application due to the limited installation space on carrier platforms. In addition, the cost of implementation and the difficulty of processing are also barriers to the application of binocular cameras in underwater scenes. By contrast, the use of monocular cameras for depth estimation tends to be less restrictive in terms of space, cost and difficulty. Nonetheless, due to the lack of reliable depth clues, such as stereo vision, monocular images can only provide relative depth (Godard et al., 2019). Fortunately, motion information in monocular sequences has been proved to be a reliable depth clue, which can be used to estimate absolute depth (Watson et al., 2021).

Monocular depth estimation can usually be realized by supervised and self-supervised methods. The supervised methods rely on a large number of labeled data. However, considering that optical properties vary from different underwater scenes, it is nearly impossible to obtain sufficient labeled data for retraining supervised methods in each new scene. Self-supervised methods, which only need video frames from different scenes, are relatively easy to deploy and quite promising in underwater scenes. The self-supervised depth estimation has greater potential for application in underwater scenes (Masoumian et al., 2022a). The main challenge of underwater self-supervised depth estimation is the lack of large-scale and available datasets like those on land. But the direct application of the depth estimation model on land to underwater scenes will be less effective. Moreover, the current underwater depth estimation is closely connected with the transmission map, which makes the obtained depth map meaningless and far away from the true depth data. Inspired by the challenges above, we propose a self-supervised model, UWdepth, able to obtain more accurate depth estimation based

* Corresponding author.
E-mail addresses: wangjunting@hrbeu.edu.cn (J. Wang), yexiufen@hrbeu.edu.cn (X. Ye).

https://doi.org/10.1016/j.engappai.2023.105846
Received 6 October 2022; Received in revised form 4 December 2022; Accepted 9 January 2023
Available online 18 January 2023

on smaller underwater datasets without GT on multiple evaluation indexes. And we can acquire the absolute depth data based on the monocular sequences, which brings wider application to our study.

UWdepth is mainly composed of iterative pose network, depth estimation network and adaptive cost volume module. The iterative pose network calculates the pose transformation matrix between two adjacent frames to reconstruct the target frame. It transforms the task of depth estimation into self-supervised image reconstruction. The adaptive cost volume module can output the geometric compatibility between successive frames at different depths, take full advantage of multi views and assist the depth estimation network to learn the mapping relationship between image pixel values and depth data. After all, we integrate the depth estimated by UWdepth with Akkaynak–Treibitz imaging model (Akkaynak and Treibitz, 2018) and further perform an image enhancement application.

The main contributions of this paper are summarized as follows:

- We propose a self-supervised model UWdepth for underwater scenes, using monocular sequences for depth estimation.
- An iterative pose network is proposed to solve the problem of complicated underwater motions and insufficient training data in underwater scenes.
- A depth consistency loss is introduced to constrain the training process, so as to achieve more accurate inter-frame motion prediction and depth estimation.
- Integrating the predicted depth and Akkaynak–Treibitz imaging model, the revised image enhancement method can provide improved visual perception for underwater images.

2. Related work

2.1. Underwater monocular depth estimation

The past study of underwater depth estimation has been inextricably linked to underwater image enhancement. The initial research was conducted by using dark channel prior (DCP) (He et al., 2010) for approximate underwater depth estimation and image recovery. Carlevaris-Bianco et al. (2010) proposed a simple but effective prior to estimate depth by using the strong difference of attenuation of three image color channels in water. Peng and Cosman (2017) proposed a depth estimation method based on image blurring and light absorption. Peng et al. (2018) estimated the ambient light by the depth-dependent color changes. Then, the transmission map can be estimated by calculating the difference between the observed intensity and ambient light. Zhou et al. (2022) developed a restoration method based on backscattered pixel prior and color cast removal, which can estimate depth map, backscattered map, and illuminant map simultaneously. All these traditional underwater depth estimation methods above are based on transmission maps. The depth results are obtained from the inverse relationship between the depth map and transmission map. Since the scattering parameters are unknown, the estimated depth maps are most likely wrong and may be far away from the depth information in a real underwater scene.

Recently, researchers have tried to estimate the depth of underwater scenes by deep-learning based methods. Due to the scarcity of datasets for underwater depth estimation, Gupta and Mitra (2019) proposed an unsupervised depth estimation method based on GAN networks, which estimated the desired depth map by learning the mapping function between unpaired RGB-D land images and arbitrary underwater images. However, this method is limited to specific water types and lacks generality. Hambarde et al. (2021) proposed a method of underwater depth image synthesis, which uses the underwater coarse-level generation network(UWC-Net) to estimate the coarse-level depth map firstly. Then, the underwater fine-level network(UWF-Net) is used to compute the fine-level depth map. However, its application will also be greatly limited due to the lack of absolute depth information clues, since it is based on single image. Based on the limitations above, our study is developed on the relatively mature monocular self-supervised depth estimation research on land. We improve it according to the characteristics of underwater scenes.

2.2. Self-supervised monocular depth estimation

Self-supervised depth estimation has attracted extensive attention since it removes the demand for GT. The basic idea of these methods is to transform the depth estimation task into an image reconstruction problem. This type of network usually depends on multiple images with specific connections, such as stereo pairs or monocular video sequences (Masoumian et al., 2022b).

Along this idea, Garg et al. (2016) proposed the first unsupervised framework to train a deep convolutional neural network for monocular depth prediction. Zhou et al. (2017) designed two independent networks to estimate depth map and camera motion in monocular video, and the model is mainly trained by reprojection loss. This pioneered monocular self-supervised depth estimation for monocular sequences. Many subsequent variants attempted to improve self-supervised monocular depth estimation. Godard et al. (2019) revised the unsupervised network framework based on binocular stereo vision (Godard et al., 2017). The pose estimation network is added to the original framework, and the original disparity estimation network is modified to direct depth estimation.

Many follow-up methods then tried to improve the self-supervised methods by the new loss terms. Bian et al. (2019) proposed the geometry consistency loss for scale-consistent predictions and an induced self-discovered mask for dealing with moving objects and occlusions. Guizilini et al. (2020) proposed a new structure from motion(SfM) framework and propose a new loss function term, speed supervised loss. Spencer et al. (2020) proposed DeFeat-Net to estimate depth by learning a cross-domain dense feature representation. And it proposed feature distortion loss and contrast loss associated with dense feature maps. Recent researches in self-supervised monocular depth estimation have been devoted to investigating occlusions in the images (Zhang et al., 2019), objects in motion between adjacent frames (Yue et al., 2022; Tosi et al., 2020) and more robust loss functions (Nakamura et al., 2021).

In this section, we will focus on depth estimation using monocular sequences, which is most promising in application. To better process the sequences of frames, C.S. Kumar et al. (2018) harnessed the ability of long short-term memory (LSTM)-based Recurrent Neural Networks (RNNs) to reason sequentially and predict the depth map for an image frame. Wang et al. (2019) proposed a dense depth map and odometry estimation method that utilized RNNs and reprojection of multi-view images. Zhang et al. (2019) proposed a novel spatial–temporal CLSTM structure, which can capture the spatial features and the temporal correlations among consecutive video frames with negligible increase in computational cost. It applies the generative adversarial learning scheme and designs a temporal consistency loss. Patil et al. (2020) produced a time series of depth maps, and integrated the corresponding networks with ConvLSTM, so that more accurate depth estimation can be generated by using the spatial–temporal structures of depth across frames. However, these methods involve extracting features from multiple frames in the sequences during training but the geometric information between frames is not employed, causing more computational demand.

Kendall et al. (2017) showed that integrating the cost volume can significantly improve results. Wimbauer et al. (2021) learned depth from the cost volume, but its training still relies on stereo pairs and sparse supervision. Khot et al. (2019) leveraged photometric consistency between multiple views as supervised signal for learning depth prediction, getting rid of the requirement for GT. However, their assumption of no moving objects in the scene imposed limitations on their application scenarios. Watson et al. (2021) solved the problems

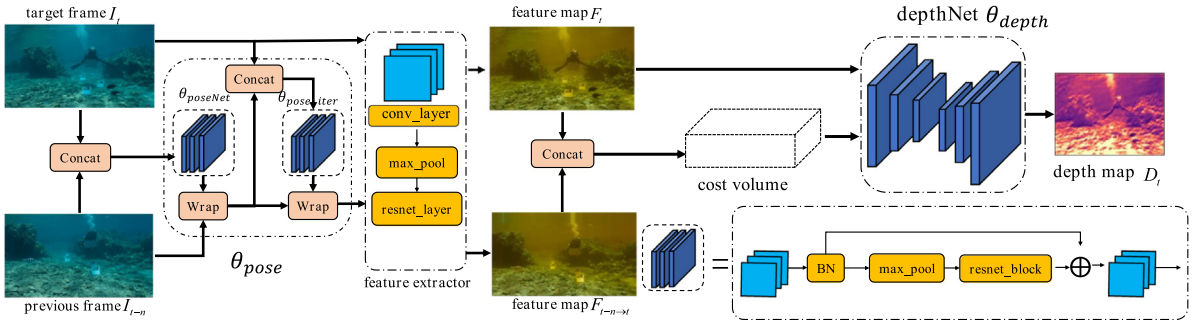


Fig. 1. The structure of UWdepth. The pose network θ_{pose} estimates the relative poses between the target frame I_t and the nearby frame I_{t-n}. It consists of two components: θ_{poseNet} and θ_{poseiter}. The warped feature map $F_{t-n\to t}$ (the result of reconstruction) and the target feature map F_t can be concatenated to build the cost volume. Then the encoder–decoder structure of the depthNet θ_{depth} will process the feature map and the cost volume to generate the depth map D_t of the target frame I_t.

associated with moving objects and static cameras and resolved the scale ambiguity in monocular self-supervised depth estimation. However, it is difficult to get satisfactory results when applying the above researches to underwater scenes. According to the characteristics of underwater scene, our research used monocular sequences for depth estimation, which makes full use of multi-perspectives based on adaptive cost volume. We proposed the iterative pose network and depth consistency loss to solve the challenges brought by more irregular underwater motion and smaller datasets in underwater scenes. With similar computational and data resources, we are able to obtain more accurate depth estimation.

3. Method

Our method combines two well organized parts: (1) A well-performing self-supervised depth estimation model UWdepth. (2) A revised imaging model for accurately simulating underwater imaging processes. Firstly, three consecutive frames are used as inputs of UWdepth and the corresponding per-pixel depth map aligned to the target frame is output. The predicted depth map is used as a parameter of the underwater imaging model to obtain the undegraded image. It constitutes a complete underwater depth estimation and image enhancement system.

In this section, we will introduce the structure of UWdepth. And we will explain the details of iterative pose network, adaptive cost volume module and our total loss function. Then the revised underwater imaging model will be presented in brief, which we use to enhance images.

3.1. Structure of UWdepth

The structure of UWdepth is shown in Fig. 1, which is composed of the iterative pose network(corresponding to θ_{poseNet} and θ_{poseiter} in Fig. 1), feature extractor, cost volume and depth estimation network (corresponding to depthNet θ_{depth} in Fig. 1). The backbone of θ_{pose} and θ_{depth} are based on ResNet18.

The purpose of depth estimation is using the trained depth estimation network θ_{depth} to output a depth map aligned with the input pixels. Firstly, we use the pose network θ_{pose} to estimate the relative pose $T_{t\to t+n}$ between the source frame I_{t+n} and the target frame I_t. Then combine $T_{t\to t+n}$ with the current estimated depth map D_t, and camera intrinsic matrix K to synthesize the scene from the same viewpoint as the target frame I_t, only using the nearby source frames I_{t+n}. It can be summarized as Eq. (1):

$$I_{t+n\to t} = I_{t+n}\left\langle \text{proj}\left(D_t, T_{t\to t+n}, K\right)\right\rangle \tag{1}$$

where $I_{t+n\to t}$ is the warped source frame from the target frame's viewpoint. proj(.) is the transformation function that maps pixels from the target image to the source image.

After processing by the feature extractor, we can get the warping feature map $F_{t+n\to t}$ and the target feature map F_t. The cost volume can be constructed by calculating the difference between $F_{t+n\to t}$ and F_t. Then we concatenate the output of cost volume with F_t, and input it to the depth estimation network to get the depth map D_t corresponding to the target frame I_t.

The loss function L_{total} can be constructed based on the above. The total loss consists of four components. The details of each loss will be introduced in Section 3.4. And we train UWdepth by minimizing the total loss.

3.2. Iterative pose network

The key point of monocular self-supervised depth estimation is synthesizing images from a new viewpoint, which requires accurate relative pose estimation. The wrong pose estimation will lead to the incorrect correspondence between the target frame and the source frame, thus leading to unreliable calculation of reprojection loss and unavoidable errors in depth prediction.

Conventionally, an independent one-step pose network is used to predict the camera's six degrees of freedom directly (Godard et al., 2019). However, the camera motion in the underwater scene is often different from that on land which mainly includes smooth translation motion. The underwater operations usually involve more rotational movement. Furthermore, underwater datasets are usually not abundant enough to train a complex pose network to predict accurate relative pose data.

To address this issue, we propose the iterative pose network, which divides the pose estimation in the underwater scene into two steps corresponding to the θ_{poseNet} and θ_{poseiter} in Fig. 1. These two modules share the same network structure but they have different training strategies and weights. θ_{poseNet} uses the initial weights pretrained on the KITTI dataset (Geiger et al., 2013). KITTI is a massive dataset, mainly applied in the autonomous driving on land. The pretrained weights of θ_{poseNet} enable it better to extract pose features from images and effectively predict regular translational motion. It is taken as fixed weights when trained to the set epoch, while θ_{poseiter} continues to be trained along with the training of θ_{depth} without pretrained weights. Based on what θ_{poseNet} has already processed, θ_{poseiter} can pay more attention to the rotational component in the motion. The θ_{poseiter} can achieve satisfactory results in relatively smaller underwater datasets. The main body of iterative pose network consists of multi-scale residual modules as shown in Fig. 1, and its structure is relatively simple and lightweight.

The pose estimation is decomposed into two steps by using the iterative pose network. Firstly, we can estimate the relative pose T_1 by θ_{poseNet}. Based on Eq. (1), we can get Eq. (2).

$$I_{t+n\to t'} = I_{t+n}\left\langle \text{proj}\left(D_t, T_1, K\right)\right\rangle \tag{2}$$

then the relative pose T_2 can be obtained by θ_{poseiter}, so we can get Eq. (3).

$$I_{t'\to t} = I_{t'} \left\langle \text{proj}\left(D_t, T_2, K\right)\right\rangle \tag{3}$$

It forms a transition state during the process of image reconstruction, creating better robustness in the process of pose estimation. Owing to the more complex motion of underwater scenes, the network in this paper can obtain a more accurate camera relative pose than that of a single step prediction based on small underwater datasets. It can establish more reliable reprojection loss and predict the depth map of the underwater image more accurately. The following experiments will strongly prove the effectiveness of the proposed iterative pose estimation network.

3.3. Adaptive cost volume

The cost volume is constructed by using multiple input frames, which is used to measure the geometric compatibility of the pixels of the target frame I_t with the nearby source frame I_{t+n} at different depth values.

Firstly, the feature extractor encode the source frame and target frame into the feature map as F_{t+n} and F_t. It assumes that the maximum and minimum depth values of this scene are $D_{\max}$ and $D_{\min}$ correspondingly. And it can define a set of ordered planes γ, with the depth of each plane linearly space between $D_{\max}$ and $D_{\min}$. For every $d \in \gamma$, we can combine it with the relative pose matrix $T_{t\to t+n}$ and the camera intrinsic matrix K to warp the feature map of the source frame F_{t+n}. We can get the warped feature map $F_{t+n\to t,d}$ based on Eq. (1). Then the cost volume is constructed as the absolute difference between the warped feature map $F_{t+n\to t,d}$ and the target feature map F_t.

Theoretically, when $d \in \gamma$ represent the correct depth value for every pixel (i,j) in the target frame I_t, the corresponding value of cost volume will be the smallest. After the activated function, the value of cost volume is converted into the probability value. As stated above, for every pixel (i,j) in the target frame, the cost volume can effectively represent the probability that the correct depth value is d as $d \in \gamma$.

Then, the output of the cost volume is connected with the feature map F_t, and then input it to the depth estimation network, which can train the network to learn the distribution of the data and the mapping relationship between the cost value and the depth data.

To solve the problem of scale ambiguity in monocular depth estimation. It learns $D_{\max}$ and $D_{\min}$ in the training process. In the training process, the maximum and minimum values of the predicted depth are calculated for each batch of training. They are used as momentum adjusting coefficients to adjust $D_{\max}$ and $D_{\min}$. After training, $D_{\max}$ and $D_{\min}$ will constantly approach the depth value of the real world and it gradually narrows down. These two parameters are constantly adjusted and saved as parameters together with model weights.

3.4. Loss function

The total loss function includes four parts, the reprojection loss L_{pe}, the smoothness loss L_s, the depth consistency loss L_{cd} and the consistency loss L_c.

Inspired by Bian et al. (2019), the result of depth prediction of reconstructed image $I_{t+n\to t}$ should be the same as the depth estimation of the target image I_t. So we add an additional depth consistency loss L_{cd}. The depth estimation generated from the preceding and following frames in sequences should be depth consistent. Specifically, we use the warped feature map $F_{t+n\to t}$ to predict the corresponding depth map $D_{t+n\to t}$. We can compute L_{cd} as shown in Eq. (4).

$$L_{cd} = \frac{\left|\overline{D}_{t+n\to t} - \overline{D}_t\right|}{\overline{D}_{t+n\to t} + \overline{D}_t} \cdot \frac{\left|D_{t+n\to t} + D_t\right|}{D_{max} + D_{min}} \tag{4}$$

where $\overline{D}_{t+n\to t}$ and $\overline{D}_t$ are the results of normalized depth estimation. D_{max} and D_{min} correspond to the maximum and minimum depth estimation value obtained from the adaptive cost volume above. L_{cd} can be adaptively weighted according to different depth values. The weight of this loss becomes larger with the increase of depth values, thus allowing faster adjustment of the larger depth values and errors, and finer adjustment of the smaller depth values. The model predicts and converges to more correct depth data with such constraint.

For the reprojection loss L_{pe}, the target frame I_t is synthesized from different nearby source frames I_{t+n}. Following Godard et al. (2019), we can calculate the minimum L_{pe} of every pixel as Eq. (5), where N is the total number of frames participating in the training.

$$L_{pe} = \min_N \left(pe\left(I_{t+n}, I_t\right)\right) \tag{5}$$

The reprojection loss L_{pe} is a linear weighted combination between L_1 loss and L_{ssim} loss which represents the structural similarity loss. We can summarize it as Eq. (6). We set $\alpha = 0.15$ here.

$$pe\left(I_{t+n}, I_t\right) = \frac{\alpha}{2} * L_1\left(I_{t+n\to t}, I_t\right) + (1-\alpha) * \cdot L_{ssim}\left(I_{t+n\to t}, I_t\right) \tag{6}$$

We can calculate L_s as shown in Eq. (7).

$$L_s = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|} \tag{7}$$

where d_t^* is the average normalized depth. L_s is used to prevent the shrinkage of the gradient. The smoothness loss L_s smooths the depth map by adding L_1 constraint to the gradient of the depth map, so as to avoid discontinuities in the depth map where the gradient changes too much (Godard et al., 2017).

In order to avoid the cost volume module showing the opposite effect in some extreme cases. Following Watson et al. (2021), we can construct the consistency loss L_c and generate a mask M that represents the construction of cost volume in the image is reliable pixel area. The final loss is as Eq. (8). We set $\tau = 0.001$ here.

$$L_{\text{total}} = (1-M)\left(L_{pe} + L_{cd} + L_c\right) + \tau * L_s \tag{8}$$

3.5. Underwater imaging model

The underwater imaging model can be expressed as Eq. (9):

$$I_c = D_c + B_c \tag{9}$$

where c=r,g,b is the color channel, I_c is the image captured by the camera and needed to be enhanced, D_c is the attenuated signal and B_c is the backscatter signal. The Akkaynak–Treibitz imaging model firstly points out that D_c and B_c are governed by two different coefficients, which were often considered equal in previous studies. The Akkaynak–Treibitz imaging model can be expressed as Eq. (10).

$$I_c = J_c e^{-\beta_c^D(V_D)Z} + B_c^{\infty}\left(1 - e^{-\beta_c^B(V_B)Z}\right) \tag{10}$$

where β_c^B and β_c^D are backscatter coefficient and light attenuation coefficient, respectively. J_c is the unattenuated scene to be recovered. Z is the distance between the camera and the scene along the line of sight. B_c^{∞} is the veiling light. Based on Eq. (10), β_c^B, β_c^D, B_c^{∞} and Z need to be estimated to recover J_c. In other words, the restoration of underwater images is simplified to the solution problem of these four parameters. Following Akkaynak and Treibitz (2019), we can estimate β_c^B, β_c^D and B_c^{∞}, so we can recover J_c by Eq. (11).

$$J_c = \left(I_c - B_c^{\infty}\left(1 - e^{-\beta_c^B(V_B)Z}\right)\right) * e^{\beta_c^D(V_D)Z} \tag{11}$$

Based on the above, Z is the only unknown parameter. We can simplify the problem of image enhancement to depth estimation. Using the depth information obtained above, we are able to perform effective underwater image enhancement.

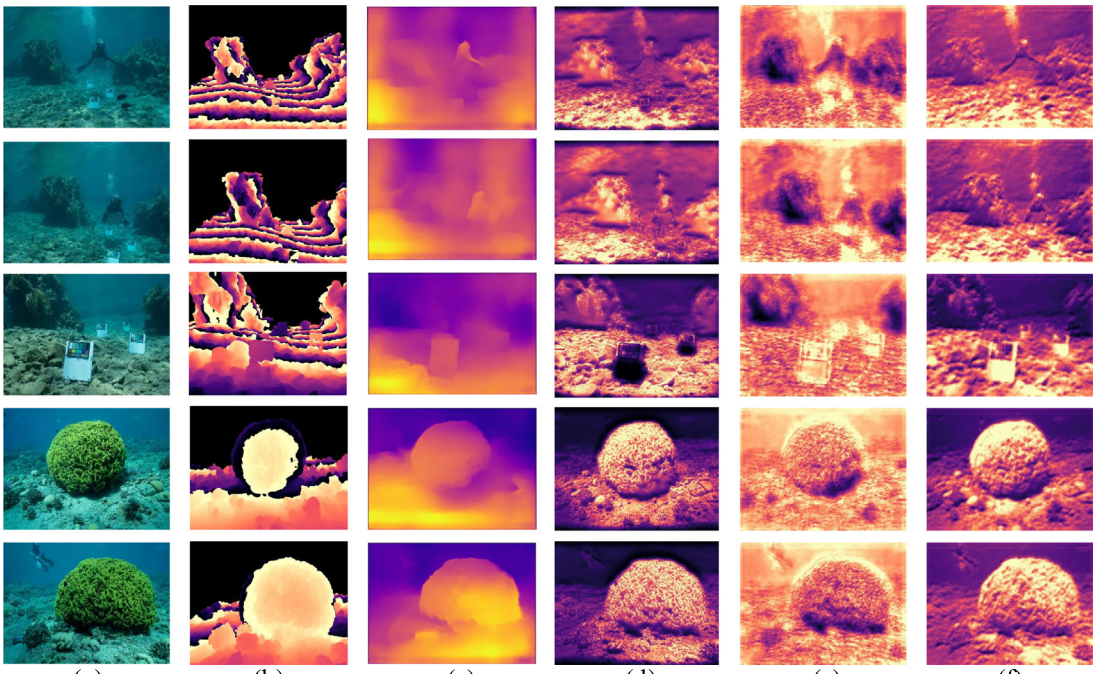


Fig. 2. The visualization of different depth estimation algorithms. (a) Raw. (b) GT. (c) densedepth (Alhashim and Wonka, 2018). (d) monodepth2 (Godard et al., 2019). (e) manydepth (Watson et al., 2021). (f) UWdepth (ours).

4. Experiments and discussion

In this section, we evaluate the effectiveness of our algorithm as follows: (1) We have done qualitative and quantitative comparative experiments to verify the superior performance of UWdepth. (2) We demonstrate the promising quality of our enhancement results in comparison with a range of conventional and deep-learning based image enhancement algorithms. (3) We have verified that each components we proposed has a meaningful effect on the improvement of overall performance through a series of ablation experiments.

In these experiments, we use the dataset from Akkaynak and Treibitz (2019), the standard depth evaluation metrics are used according to Eigen and Fergus (2015), Eigen et al. (2014). Meanwhile, MSE, PSNR and SSIM (full-reference evaluation), UCIQE and UIQM (no-reference evaluation) are used as the evaluation indexes of image enhancement.

4.1. Implementation details

Our network was implemented by the Pytorch framework. The model was trained and tested on a NVIDIA GeForce RTX 3060 Ti GPU. The total number of parameters of our whole model is 34.38M, which is of medium size. We utilized the Adam optimizer to optimize the model with a batch size of 2. We set the learning rate as 0.0001. We optimize a model with 50 epochs, and set the freeze_epoch(epoch of θ_{poseNet} training) as 25.

The whole dataset consists of 1111 images. We divide 1000 images into the training set and 111 images into the testing set. The original resolution of these images are 7360*4912 or 7952*5309. In order to reduce the demand for computing resources, we process the images to a uniform resolution of 640 * 480.

4.2. Contrastive experiment for depth estimation

We compare UWdepth with densedepth (Alhashim and Wonka, 2018), monodepth2 (Godard et al., 2019) and manydepth (Watson et al., 2021). The results of monodepth2 (Godard et al., 2019) and manydepth (Watson et al., 2021) have been trained on the same dataset we used, considering these two methods and our models are self-supervised methods. Fig. 2 visualizes the results of depth maps predicted by different depth estimation models. As we can see, the depth map predicted by UWdepth is more refined compared to other models. Our depth estimation results are closer to the GT in the rough contour of estimation. As we can see, the depth estimation results of several self-supervised methods all contain more texture information. We can reasonably guess that the depth range of our underwater dataset is smaller than that on land and the depth value of the same target will also change. The depth at the same target in GT also varies greatly, unlike the same target on land which is a smooth plane. It is difficult to objectively evaluate the depth estimation results from qualitative analysis alone.

Table 1 shows the quantitative comparison results of different depth estimation models. With the decreasing of absolute difference (AbsRel), root mean square error (Rmse), Rmse(Log) and relative error SqRel, the results of depth estimation get closer to GT. a_1, a_2, a_3 are the percentages of all pixels that satisfy the Eq. (12).

$$\max\left(\frac{d_i}{d_{GT}}, \frac{d_{GT}}{d_i}\right) < thr \tag{12}$$

where d_i is the depth value we predict, d_{GT} is the GT of the depth data, thr is threshold which $thr = 1.5$, $thr = 1.5^2$, $thr = 1.5^2$ here. A larger value means a more accurate result.

We use boldface to indicate the best result in each index. Specifically, compared to the recent self-supervised model manydepth, our method reduces AbsRel by 0.143 and increases a_1 by 0.275, reaching an AbsRel of 0.190 and a_1 of 0.685. Furthermore, even compared to the supervised model densedepth, we observe an improved performance from 0.354 down to 0.190 for AbsRel and from 0.384 up to 0.685 for a_1. In addition, the ability of absolute depth obtained by UWdepth brings more extensive application for subsequent development.

4.3. Contrastive experiment for image enhancement

We have compared our algorithm with various traditional underwater image enhancement algorithms (UDCP (Drews et al., 2013),

Table 1
The quantitative comparison of different depth estimation algorithms.

Method	Self Supervised	Absolute Depth Data	abs_rel↓	sq_rel↓	Rmse↓	Rmse_log↓	a1↑	a2↑	a3↑
Densedepth (Alhashim and Wonka, 2018)	×	✓	0.354	0.774	1.903	0.404	0.384	0.704	0.888
Monodepth2 (Godard et al., 2019)	✓	×	0.242	0.384	1.431	**0.298**	0.519	**0.875**	0.977
Manydepth (Watson et al., 2021)	✓	✓	0.333	0.705	1.803	0.376	0.410	0.734	0.918
UWdepth	✓	✓	**0.190**	**0.097**	**0.415**	0.300	**0.685**	0.857	**0.988**

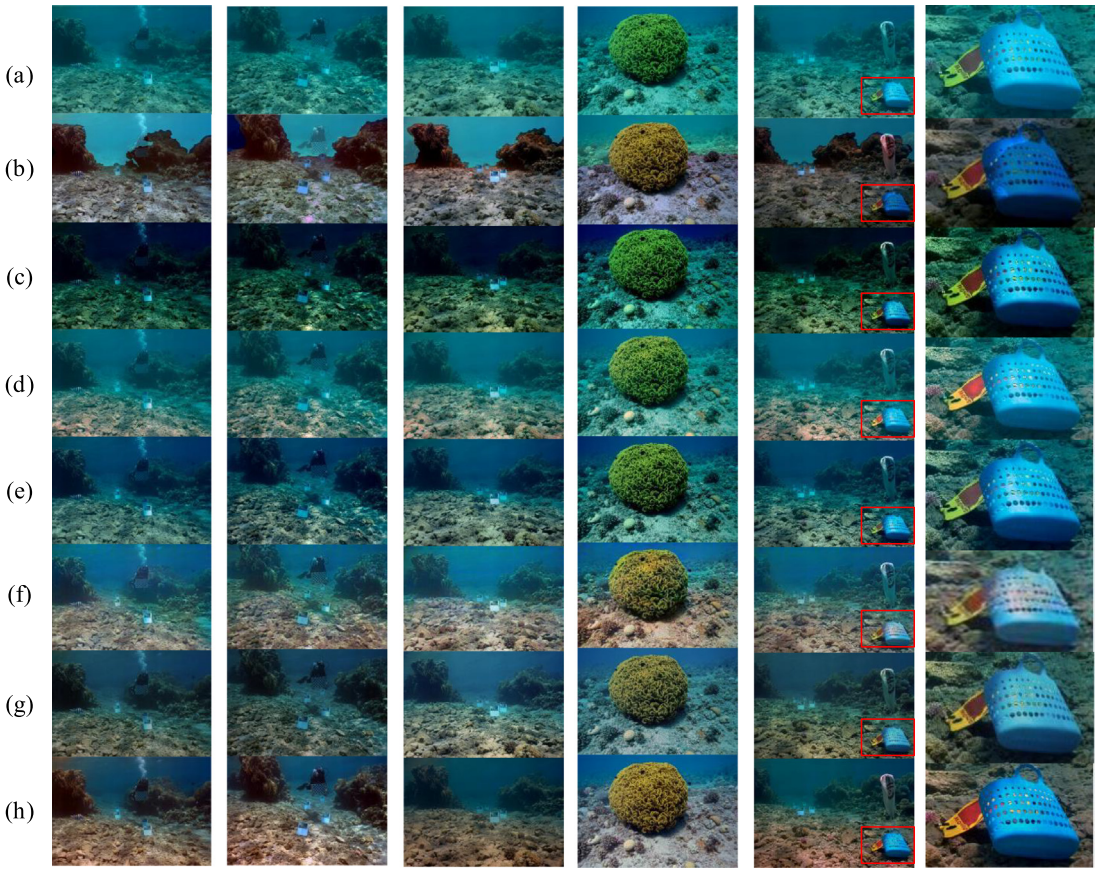


Fig. 3. The qualitative comparison of different image enhancement algorithms. (a) Raw. (b) GT. (c) UDCP (Drews et al., 2013). (d) IBLA (Peng and Cosman, 2017). (e) ULAP (Song et al., 2018). (f) Funie-GAN (Islam et al., 2020). (g) WaterNet (Li et al., 2019). (h) UWdepth (ours). The last column is an enlarged view of the area corresponding to the red box.

IBLA (Peng and Cosman, 2017), ULAP (Song et al., 2018)) and deep-learning based image enhancement algorithms (Funie-GAN (Islam et al., 2020), Water-Net (Li et al., 2019)). Fig. 3 gives the qualitative comparison of different underwater image enhancement methods. As presented, the raw underwater images show blue–green tones and low illumination. It is clear that the images enhanced by our method perform better in color correction and blur elimination. In the last column, we added an enlarged view of the area with red box. The enlarged image obtained by our method is obviously better in structural recovery and color naturalness. Compared with GT, our method produces the closest results, which proves our method's remarkable performance in quality improvement.

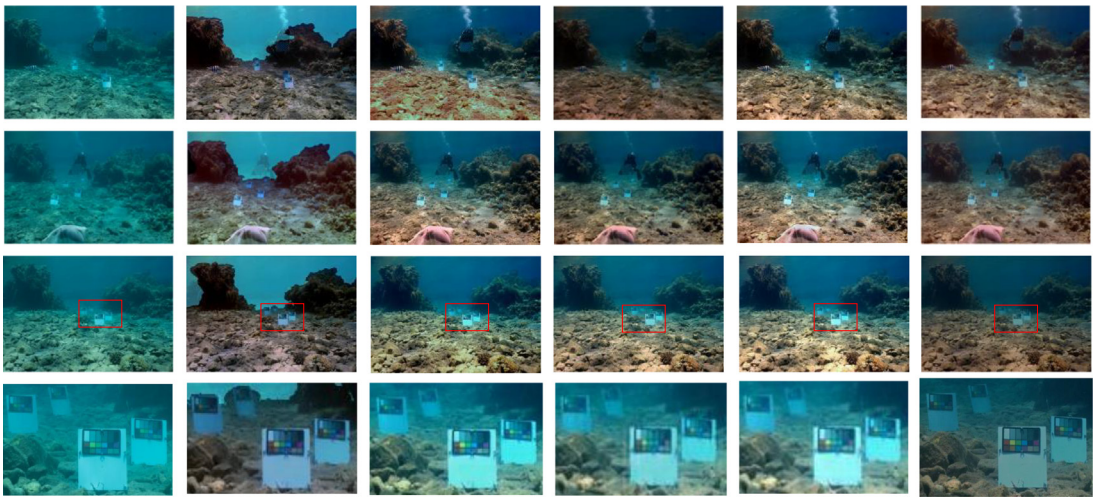
To quantitatively evaluate the performance of different methods, we have carried out the full-reference evaluation and non-reference evaluation. As shown in Table 2, we use three indexes (MSE, PSNR and SSIM) to conduct a comprehensive full-reference evaluation. Higher PSNR scores and lower MSE scores indicate that the results are closer to the reference image in image content, while higher SSIM scores indicate that the result is closer to the reference image in image structure and texture.

We use boldface to indicate the best result in each index. It shows the corresponding improvement of 42.6%, 35.4%, 40.0%, 24.3% and 20.6% over the UDCP (Drews et al., 2013), IBLA (Peng and Cosman, 2017), ULAP (Song et al., 2018), Funie-GAN (Islam et al., 2020) and Water-Net (Li et al., 2019) in terms of pnsr. Meanwhile, compared with the algorithms of the corresponding order above, our method improves SSIM by 75.8%, 50.9%, 54.1%, 48.3% and 30.3% respectively. There is no doubt that our method achieves the best results in all three full-reference indexes, which is far superior to the second best method.

Table 2
Full-Reference evaluation.

Method	MSE(*1000)↓	PSNR↑	SSIM↑
UDCP (Drews et al., 2013)	3.6577	15.384	0.5155
IBLA (Peng and Cosman, 2017)	1.7158	16.197	0.6005
ULAP (Song et al., 2018)	1.9416	15.669	0.5879
Funie-GAN (Islam et al., 2020)	1.2320	17.658	0.6109
WaterNet (Li et al., 2019)	1.1909	18.187	0.6955
UWdepth	**0.5145**	**21.941**	**0.9062**

We present the non-reference evaluation in Table 3. The higher the UCIQE score is, the better saturation and contrast of results will be. The higher the UIQM scores, the more the results are in line with human visual perception. We can observe that traditional methods tend



(a) (b) (c) (d) (e) (f)

Fig. 4. Visualization of the results of depth estimation ablation experiments. (a) Raw. (b) GT. (c) Baseline (Watson et al., 2021). (d) Baseline + L_{cd}. (e) Baseline + θ_{poseiter}. (f) UWdepth. The last line is an enlarged view of the area corresponding to the red box.

Table 3
No-Reference evaluation.

Method	UCIQE↑	UIQM↑
UDCP (Drews et al., 2013)	**0.5147**	**1.7907**
IBLA (Peng and Cosman, 2017)	0.4722	1.4731
ULAP (Song et al., 2018)	0.4844	1.6026
Funie_GAN (Islam et al., 2020)	0.4246	0.8687
WaterNet (Li et al., 2019)	0.3971	0.7290
UWdepth	0.5105	0.9706

to produce better UCIQE and UIQM scores. But compared with other methods based on deep-learning, we reach the best results. We will discuss this in detail in Section 4.5.

4.4. Ablation study

We perform ablation studies for our design of the iterative pose network and the additional depth consistency loss. As shown in Table 4, we verify the validity of each component by turning on and off these changes in turn.

Compared with baseline, the performance can be improved by using the iterative pose network, reducing the AbsRel reduce from 0.333 to 0.232 and increasing a_1 from 0.410 to 0.613. Meanwhile, by adding the depth consistency loss L_{cd}, we find that the AbsRel reduce from 0.333 to 0.272, a_1 increases from 0.410 to 0.470. Furthermore, by applying these two together, substantial improvements can be achieved in all evaluation indexes. It can be observed that the AbsRel reduces from 0.333 to 0.190, a_1 increases from 0.410 to 0.685.

Meanwhile, in order to show the different results of depth estimation more intuitively, the relevant image enhanced ablation experiments are carried out. We enhance these images based on the depth estimation results of the above ablation experiments and conduct qualitative and quantitative analysis.

Fig. 4 shows the image enhancement results of different ablation experiments. We note that the visual quality of image enhancement results is consistent with the depth estimation indexes in Table 2. The improvement of depth estimation indexes deliver greater image enhancement results correspondingly. By adding the iterative pose network and depth consistency loss, the final enhanced images demonstrate better contrast and details which produce more visually pleasing results. In the enlarged area, our final results show that the color reduction and clarity of reef and color chart are obviously the best, which is significantly improved compared with baseline. Table 5 illustrates the results of the quantitative analysis. The experimental data reinforces the usefulness of each component we propose, which is consistent with the results of the depth estimation ablation experiments.

4.5. Discussion on evaluation metric

As shown in Section 4.3, our algorithm does not produce the best UCIQE and UIQM scores. We explore it in this section, discussing whether these two indexes can reasonably represent the effect of image enhancement. As shown in Fig. 5, as far as human visual effects are concerned, it is obvious that UWdepth produces the best results, but UDCP reaches the highest score.

Therefore, we analyze the principle of UCIQE and UIQM used to evaluate the image enhancement effect. UCIQE is a linear combination of color density, saturation and contrast, which is used to quantitatively evaluate the problems of uneven color cast, blurring and low contrast of underwater images. UIQM is used to evaluate the color balance of three color channels and the contrast of the image. Traditional methods often enhance the image by improving the image contrast and solving color cast and blur, which is consistent with the direction measured by these two indexes. This is why most traditional algorithms can perform better on these non-reference indexes, but they do not conform to the visual effect.

The evaluation index is useful to a certain extent, but if it exceeds a certain limit, it is not the larger the index, the better the effect. Non-reference evaluation indicators are biased. Traditional methods are generally better than those deep-learning based methods. This encourages us to explore more logical and practical evaluation indexes of underwater images.

4.6. Application test

Referring Zhuang et al. (2022), we have demonstrated the practicability of our method for several challenging applications, without any fine tuning of parameters. Firstly, based on Yang et al. (2013), we detect the corresponding saliency of different image enhancement algorithm. As shown in Fig. 6, we can see that our saliency map can better capture texture changes in the image and produce saliency detection results closer to GT. We can obviously observe that deep-learning based methods generally outperform traditional methods in capturing texture transformations and are superior in saliency detection. Meanwhile, we adopt the SIFT keypoint detection based on Lowe (2004). As shown in Fig. 7, we can clearly see that our method can produce more keypoints at the boundary of coral reefs than other competitors. Experimental results show that our enhancement method is better at recovering the key features of underwater targets.

Table 4
The results of depth estimation ablation experiments.

Method	abs_rel↓	sq_rel↓	Rmse↓	Rmse_Log↓	a1↑	a2↑	a3↑
Baseline (Watson et al., 2021)	0.333	0.705	1.803	0.376	0.410	0.734	0.918
Baseline + L_{cd}	0.272	0.477	1.546	0.318	0.470	0.826	0.963
Baseline + θ_{poseiter}	0.232	0.122	0.454	0.341	0.613	0.834	0.922
UWdepth	**0.190**	**0.097**	**0.415**	**0.300**	**0.685**	**0.857**	**0.988**

Table 5
The results of ablation experiments with image enhancement indexes.

Method	MSE(*1000)↓	PSNR↑	SSIM↑	UCIQE↑	UIQM↑
Baseline (Watson et al., 2021)	1.946	15.384	0.694	0.499	0.627
Baseline + L_{cd}	1.882	16.010	0.729	0.503	0.739
Baseline + θ_{poseiter}	1.069	19.644	0.824	0.508	0.881
UWdepth	**0.515**	**21.941**	**0.906**	**0.511**	**0.971**

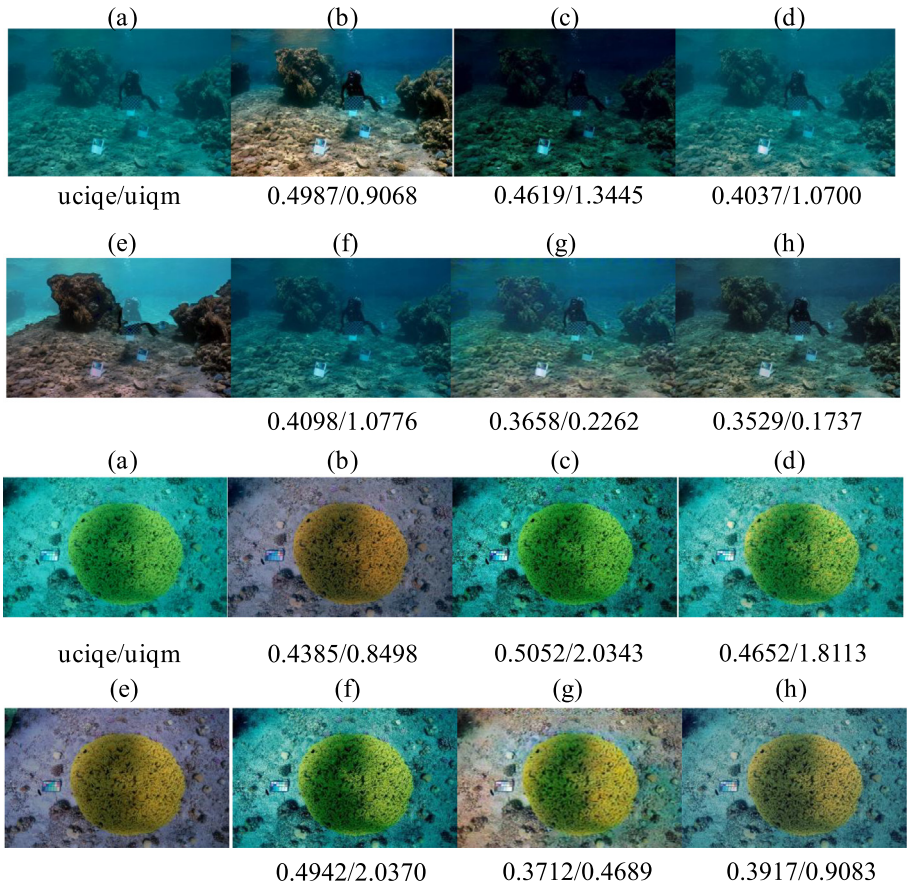


Fig. 5. The UCIQE and UIQM scores correspond to different image enhancement algorithms. (a) Raw. (b) UWdepth. (c) UDCP (Drews et al., 2013). (d) IBLA (Peng and Cosman, 2017). (e) GT. (f) ULAP (Song et al., 2018). (g) Funie_GAN (Islam et al., 2020). (h) WaterNet (Li et al., 2019).

5. Conclusion

We propose UWdepth, a model for self-supervised underwater depth estimation with monocular sequences. Considering the characteristics of underwater scenes, we propose an iterative pose network and introduce a depth consistency loss. Using the predicted depth, we can perform exact image enhancement with the Akkaynak–Treibitz imaging model. The best results of AbsRel, SqRel, Rmse can be generated on the sea-thru dataset (Akkaynak and Treibitz, 2019) by applying UWdepth. Besides, our image enhancement results can arise the best indexes for PSNR, SSIM.

Due to the size of the dataset, the accuracy of the depth estimation in this study cannot be compared with that achieved on land. There is still much room for improvement. Since it does not take into account the different optical properties of underwater scenes, the model will lack a certain level of generality. It is one of our next research, taking account of the underwater optical characteristics to enhance the generalization performance of the model. Considering the future work, the first thing to be solved is the dataset. Effective prior knowledge can reduce the demand for data resources to a certain extent. Meanwhile, different from the depth estimation on land, the depth information of underwater images is mainly used for image enhancement or restoration currently. We will try to integrate the downstream tasks and make

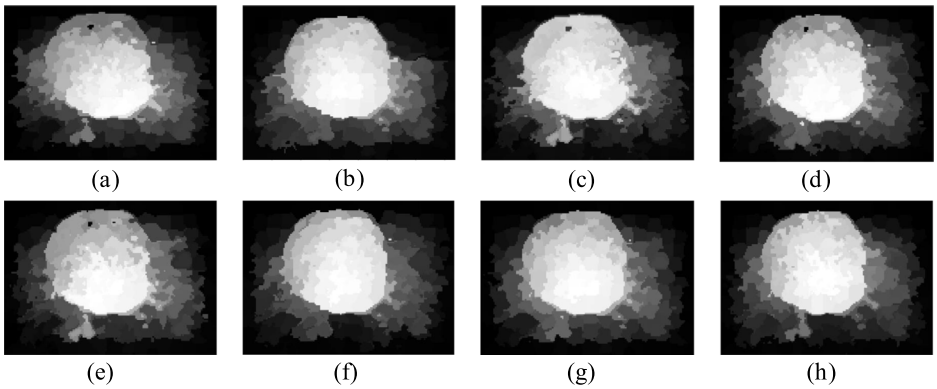


Fig. 6. The application of saliency detection. (a) Raw. (b) GT. (c) UDCP (Drews et al., 2013). (d) IBLA (Peng and Cosman, 2017). (e) ULAP (Song et al., 2018). (f)Funie_GAN (Islam et al., 2020). (g) WaterNet (Li et al., 2019). (h) UWdepth (ours).

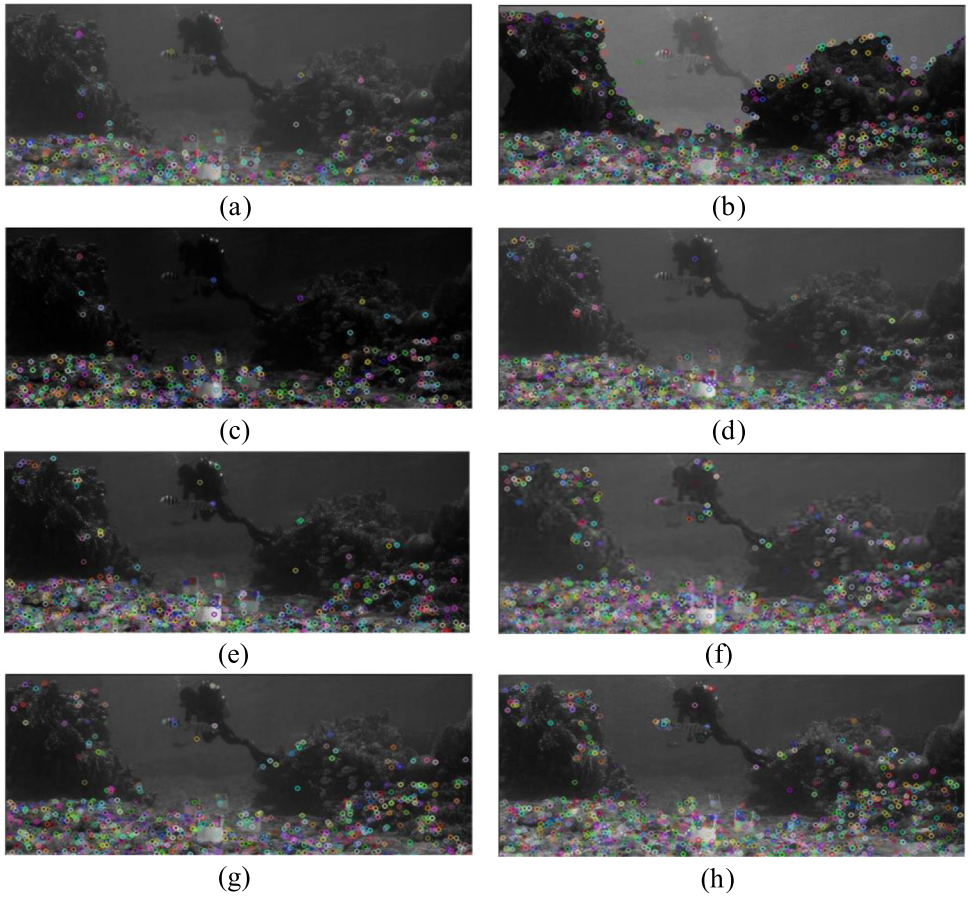


Fig. 7. The application of keypoint detection. (a) Raw. (b) GT. (c) UDCP (Drews et al., 2013). (d) IBLA (Peng and Cosman, 2017). (e) ULAP (Song et al., 2018). (f) Funie_GAN (Islam et al., 2020). (g) WaterNet (Li et al., 2019). (h) UWdepth (ours).

full use of the correlation between these two tasks to improve their performance correspondingly.

CRediT authorship contribution statement

Junting Wang: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Xiufen Ye:** Conceptualization, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Yusong Liu:** Conceptualization, Writing – review & editing, Methodology. **Xinkui Mei:** Resources, Investigation, Visualization. **Jun Hou:** Investigation, Resources, Data curation.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 42276187 and 41876100) and the Fundamental Research Funds for the Central Universities, China (Grant No. 3072022FSC0401).

References

Akkaynak, D., Treibitz, T., 2018. A revised underwater image formation model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6723–6732.

Akkaynak, D., Treibitz, T., 2019. Sea-thru: A method for removing water from underwater images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1682–1691.

Alhashim, I., Wonka, P., 2018. High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941.

Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.-M., Reid, I., 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. Adv. Neural Inf. Process. Syst. 32.

Carlevaris-Bianco, N., Mohan, A., Eustice, R.M., 2010. Initial results in underwater single image dehazing. In: Oceans 2010 Mts/IEEE Seattle. IEEE, pp. 1–8.

C.S. Kumar, A., Bhandarkar, S.M., Prasad, M., 2018. Depthnet: A recurrent neural network architecture for monocular depth prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 283–291.

Drews, P., Nascimento, E., Moraes, F., Botelho, S., Campos, M., 2013. Transmission estimation in underwater single images. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 825–830.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2650–2658.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Adv. Neural Inf. Process. Syst. 27.

Garg, R., Bg, V.K., Carneiro, G., Reid, I., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision. Springer, pp. 740–756.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. Int. J. Robot. Res. 32 (11), 1231–1237.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 270–279.

Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J., 2019. Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838.

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A., 2020. 3D packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2485–2494.

Gupta, H., Mitra, K., 2019. Unsupervised single image underwater depth estimation. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 624–628.

Hambarde, P., Murala, S., Dhall, A., 2021. UW-GAN: Single-image depth estimation and image enhancement for underwater images. IEEE Trans. Instrum. Meas. 70, 1–12.

He, K., Sun, J., Tang, X., 2010. Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. 33 (12), 2341–2353.

Islam, M.J., Xia, Y., Sattar, J., 2020. Fast underwater image enhancement for improved visual perception. IEEE Robot. Autom. Lett. 5 (2), 3227–3234.

Jian, M., Liu, X., Luo, H., Lu, X., Yu, H., Dong, J., 2021. Underwater image processing and analysis: A review. Signal Process., Image Commun. 91, 116088.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 66–75.

Khot, T., Agrawal, S., Tulsiani, S., Mertz, C., Lucey, S., Hebert, M., 2019. Learning unsupervised multi-view stereopsis via robust photometric consistency. arXiv preprint arXiv:1905.02706.

Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D., 2019. An underwater image enhancement benchmark dataset and beyond. IEEE Trans. Image Process. 29, 4376–4389.

Li, X., Hou, G., Li, K., Pan, Z., 2022. Enhancing underwater image via adaptive color and contrast enhancement, and denoising. Eng. Appl. Artif. Intell. 111, 104759.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Masoumian, A., Rashwan, H.A., Abdulwahab, S., Cristiano, J., Asif, M.S., Puig, D., 2022a. Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network. Neurocomputing.

Masoumian, A., Rashwan, H.A., Cristiano, J., Asif, M.S., Puig, D., 2022b. Monocular depth estimation using deep learning: A review. Sensors 22 (14), 5353.

Nakamura, A.T.M., Grassi Jr., V., Wolf, D.F., 2021. An effective combination of loss gradients for multi-task learning applied on instance segmentation and depth estimation. Eng. Appl. Artif. Intell. 100, 104205.

Patil, V., Van Gansbeke, W., Dai, D., Van Gool, L., 2020. Don't forget the past: Recurrent depth estimation from monocular video. IEEE Robot. Autom. Lett. 5 (4), 6813–6820.

Peng, Y.-T., Cao, K., Cosman, P.C., 2018. Generalization of the dark channel prior for single image restoration. IEEE Trans. Image Process. 27 (6), 2856–2868.

Peng, Y.-T., Cosman, P.C., 2017. Underwater image restoration based on image blurriness and light absorption. IEEE Trans. Image Process. 26 (4), 1579–1594.

Raveendran, S., Patil, M.D., Birajdar, G.K., 2021. Underwater image enhancement: a comprehensive review, recent trends, challenges and applications. Artif. Intell. Rev. 54 (7), 5413–5467.

Song, W., Wang, Y., Huang, D., Tjondronegoro, D., 2018. A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In: Pacific Rim Conference on Multimedia. Springer, pp. 678–688.

Spencer, J., Bowden, R., Hadfield, S., 2020. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14402–14413.

Tosi, F., Aleotti, F., Ramirez, P.Z., Poggi, M., Salti, S., Stefano, L.D., Mattoccia, S., 2020. Distilled semantics for comprehensive scene understanding from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4654–4665.

Wang, R., Pizer, S.M., Frahm, J.-M., 2019. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5555–5564.

Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M., 2021. The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1164–1174.

Wimbauer, F., Yang, N., Von Stumberg, L., Zeller, N., Cremers, D., 2021. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6112–6122.

Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.-H., 2013. Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3166–3173.

Yue, M., Fu, G., Wu, M., Zhang, X., Gu, H., 2022. Self-supervised monocular depth estimation in dynamic scenes with moving instance loss. Eng. Appl. Artif. Intell. 112, 104862.

Zhang, H., Shen, C., Li, Y., Cao, Y., Liu, Y., Yan, Y., 2019. Exploiting temporal consistency for real-time video depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1725–1734.

Zhou, T., Brown, M., Snavely, N., Lowe, D.G., 2017. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858.

Zhou, J., Yang, T., Chu, W., Zhang, W., 2022. Underwater image restoration via backscatter pixel prior and color compensation. Eng. Appl. Artif. Intell. 111, 104785.

Zhuang, P., Wu, J., Porikli, F., Li, C., 2022. Underwater image enhancement with hyper-Laplacian reflectance priors. IEEE Trans. Image Process. 31, 5442–5455.